Numerical Weather Prediction

Verification of the fit of rainfall analyses and forecasts to radar.

This is the stage 5 report from the Storm Scale Modelling project.
The project is half funded by Defra, project code FD2207

Forecasting Research Technical Report No. 442
Joint Centre for Mesoscale Meteorology Report No. 148

Nigel Roberts

©Crown Copyright

Verification of the fit of rainfall analyses and forecasts to radar. Stage 5 report from the Storm-Scale numerical modelling project

PROJECT DOCUMENTATION

PRODUCT DESCRIPTIONS

First Release:
Date: 31st March 2004
Latest release:
Date: 24th May 2004

PRINCE2

Author: Nigel Roberts

Owner: Nigel Roberts

Client:

Document Number: Version 2.0

Product Descriptions History

Revision History

Date of this revision: 24/05/04
Date of Next revision:

Revision date	Previous revision date	Summary of Changes	Changes marked
	31/03/04	First issue, version 1.0	
	22/04/04	Changes following comments by Peter Clark 05/04/03, version 1.1	
	10/05/04	Further changes to tidy up and more added to appendix , version 1.2	
24/05/04		Addition of report numbers, version 2.0	

Approvals

This document requires the following approvals. Signed approval forms are filed in the Quality section of the PCB.

Name	Signature	Title	Date of Issue	Version
Brian Golding	*Brian Golding*	Hd(FR)	30/6/4	2.0

Contents

1 Acknowledgements 5
2 Introduction 5
3 The model runs 6
3.1 The five model configurations 6
3.2 The model domains 7
4 The events 8
4.1 28th August 2003 06 to 22 UTC 8
4.1.1 Subjective assessment 8
4.2 13th May 2003 06 to 22 UTC 10
4.2.1 Subjective assessment 10
4.3 25th May 2003 06 to 22 UTC 12
4.3.1 Subjective assessment 12
4.4 1st July 2003 06 to 22 UTC 14
4.4.1 Subjective assessment 14
5 The objective verification approach 16
5.1 Verification scores 17
5.1.1 Fractions skill score for comparing forecast fractions against radar fractions. 17
5.1.2 The Brier skill score 17
5.1.3 The log Odds Ratio (comparing forecast and radar fractions) 17
6 Verification scores for each event 18
6.1 Scores for 6-hour accumulations 18
6.1.1 Fractions skill scores 18
6.1.2 Brier skill scores 21
6.2 Scores for hourly accumulations 23
6.2.1 Mean accumulations 23
6.2.2 Skill scores for comparison of fractions – key results 24
7 Verification scores for all events 27
7.1 Scores for 6-hour accumulations 27
7.1.1 Skill scores for comparison of fractions and odds ratios 27
7.1.2 Brier skill scores 30
7.2 Scores for hourly accumulations 31
7.2.1 Skill scores for comparison of fractions and odds ratios 31
7.2.2 Mean accumulations 33
8 Conclusions 34
8.1 Summary of the results 34
8.2 Implications of the results 35
8.2.1 The impact of higher resolution 35
8.2.2 The impact of data assimilation 36
8.2.3 Final comments 36
9 Appendices 37
9.1 Appendix A. Data assimilation methods 37
9.1.1 3DVAR 37
9.1.2 Moisture Observation Pre-Processing System (MOPS) 38
9.1.3 Addition of lower-resolution 3DVAR increments 39
9.2 Appendix B. Verification scores 40
9.2.1 Brier skill score 40
9.2.2 Fractions skill score 42
9.2.3 (log) Odds ratio 44
10 References 45

1 Acknowledgements

I would like to acknowledge at the outset that this report could not have been written without the work and expertise of several people at the Joint Centre for Mesoscale Meteorology (JCMM).
They are :-
Humphrey Lean, who managed the High Resolution Trial Model Project (HRTM), set up default configurations for the high resolution model runs, and ran most of the model forecasts used for the verification.
Zhihong Li, who got 3DVAR data assimilation working in a model with a horizontal gridlength of 4 km.
Mark Dixon, who made it possible to add 3DVAR data assimilation increments into a model with a horizontal gridlength of 1 km.
Andrew Macallan, who ran some of the forecasts.

I would also like to thank Peter Clark for comprehensive comments on the first draft.

The work was half funded by Defra.

2 Introduction

This is the 5th interim report to be delivered from the Storm Scale Numerical Modelling project. The project was set-up to investigate the ability of a storm scale configuration of the Met Office Numerical Weather Prediction (NWP) model to predict flood-producing rainfall events. In this context a storm-scale model has a horizontal grid spacing of about 1 km and receives information at the boundaries from a model with 4-km grid spacing.

The report will show results from the objective verification of precipitation forecasts run with gridlengths of 12, 4 and 1 km for 4 convective rainfall events in Spring and Summer 2003. The simulations were part of the summer 2003 testing phase of the High Resolution Trial Model (HRTM) project (Lean 2003), in which high resolution model forecasts were run with and without data assimilation. The data assimilation methods applied were 3DVAR and MOPS for model gridlengths of 12 and 4 km. These techniques are outlined in Appendix A and in the 4th interim report (Roberts 2004) with references for further information. The approach used for the 1 km simulations was to add interpolated increments from 3DVAR at 4 km rather than attempt 3DVAR at 1 km (also outlined in Appendix A and the 4th interim report) . Comparison is made with 4 and 1 km gridlength forecasts that took their starting fields from the 12 km model and were run with no extra data assimilation being applied to the 4 or 1 km grids.

The aim of the verification was twofold: (1) to determine the impact of data assimilation on high-resolution precipitation analyses and forecasts and (2) assess the accuracy of the forecasts over different spatial scales. The verification measures are described briefly in Appendix B and in more detail in the 4th interim report.

It must be emphasised that the 4 km data assimilation used here was a first version, based on the data assimilation used at 12 km, without any optimisation for a smaller domain or higher resolution. There are known issues with 3DVAR in nested domains and with MOPS at resolutions higher than that used operationally. Research is currently underway at the Joint Centre for Mesoscale Meteorology (JCMM) to find viable solutions to these difficulties and to explore new approaches.

3 The model runs

3.1 The five model configurations

For each of the four events, five different model configurations were run.

1. 12-km gridlength with 3DVAR and MOPS data assimilation
2. 4-km gridlength with 3DVAR and MOPS data assimilation
3. 4-km gridlength starting from 12-km gridlength fields with no data assimilation at 4 km. The starting point is the 12 km forecast after 1 hour (i.e. 1. at T+1).
4. 1-km gridlength with increments added from 4-km 3DVAR
5. 1-km gridlength starting from 12-km gridlength fields withg no data assimilation at 1 km. The starting point is the 12 km forecast after 1 hour (i.e. 1. at T+1).

For each of the 5 different configurations mentioned above, 4 forecasts were run for each event. Each forecast was 7 hours long. The 7-hour forecasts started at 06, 09, 12 and 15 UTC. This is pictured in Figure 1. The simulations with data assimilation (1, 2 and 4 above) were run as a continuous cycle. This means that the data assimilation was applied to a previous 3-hour forecast (T+3) at that resolution and this provided the start (T+0) for the next forecast, and so on for each cycle. The simulations that did not have extra high-resolution data assimilation (3 and 5 above) were started from 12-km gridlength T+1 forecasts i.e. when the 12-km data assimilation had finished. They had a resolution equivalent to a 12-km gridlength at T+1 and had to 'spin up' during every forecast.

It should also be mentioned as a detail for completeness, that MOPS increments were added for 30 minutes beyond T+1, but the majority of the assimilation had been done by that time.

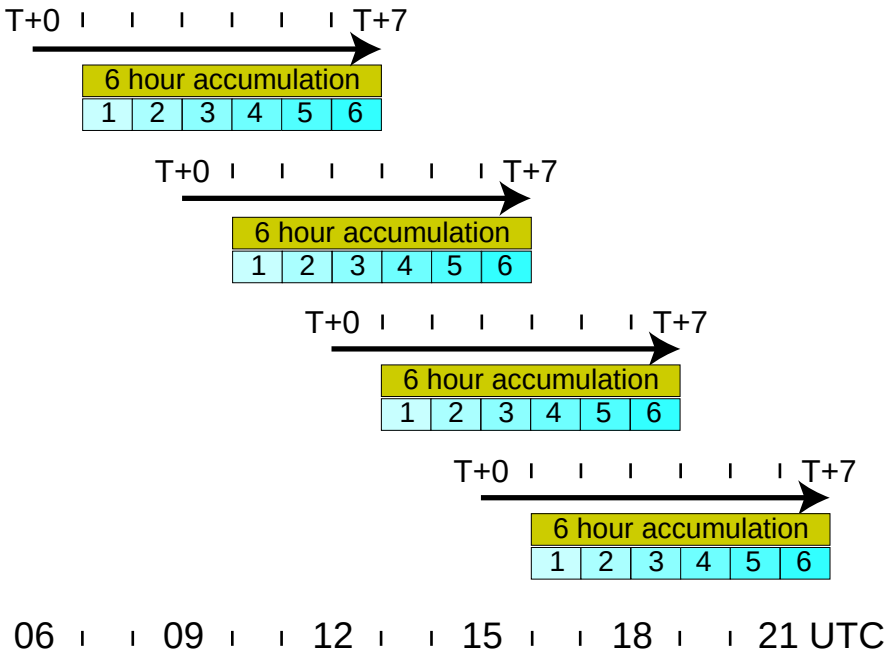


Figure 1. A schematic of the 4-cycles run for each event. The first cycle starts at 06 UTC and the last at 15 UTC. The nominal analysis time (or start of a forecast) is referred to as T+0 (06, 09, 12, 15 UTC). The forecasts end at T+7, seven hours after the start (13, 16, 19, 22 UTC). The green boxes show the verification periods for the comparison of 6-hour rainfall accumulation with radar. They were combined to give scores for each case. The blue boxes numbered 1 to 6 show the hourly accumulation periods that were verified against radar. They were also combined to give scores for each case.

Verification statistics were generated by comparing hourly and 6-hourly precipitation accumulations from T+1 to T+7 with the equivalent accumulations from the UK radar network. Figure 1 shows a schematic of the accumulation periods for each of the 4 events.

For each of the 5 forecast configurations, the statistics from the four cycles in each event were combined. Statistics from all 4 events were also combined.

The verification approach is discussed in section 5.

3.2 The model domains

Figure 2 shows the domains that were used for the 12, 4 and 1-km gridlength forecasts. The 12-km gridlength forecasts supplied the information required at the boundaries for the 4-km forecasts, which in turn, supplied information to the 1-km boundaries. All the verification scores were generated over the 1-km gridlength domain. The dimensions of the 1-km domain are 300x300 km.

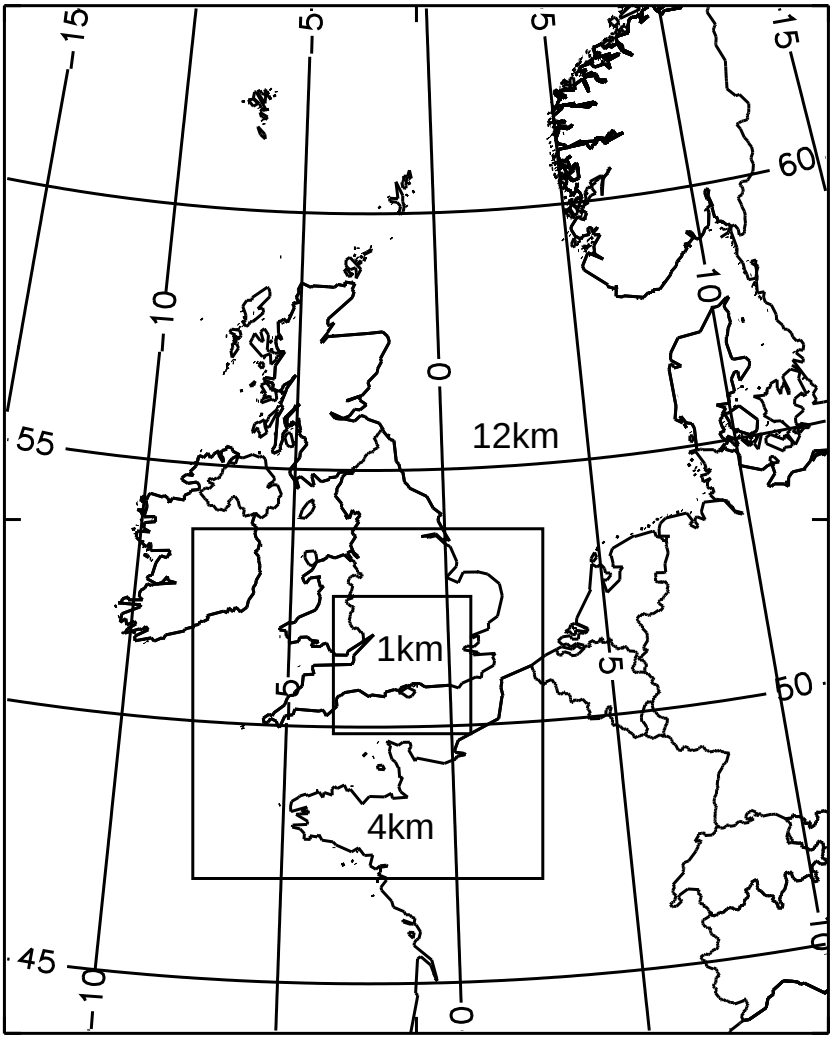


Figure 2. The domains used for the 12, 4 and 1-km gridlength simulations.

4 The events

The four events cover the periods

1. 06 to 22 UTC on 28th August 2003
2. 06 to 22 UTC on 13th May 2003
3. 06 to 22 UTC on 25th May 2003
4. 06 to 22 UTC on 1st July 2003

Each will now be described briefly along with a picture of the precipitation accumulations from radar and from the 5 model runs over the 6-hour verification period of forecast 2. This covers the period 10 to 16 UTC of each event.

4.1 28th August 2003 06 to 22 UTC

Early in the period heavy showers and thunderstorms developed over the English Channel and advected across the southeast of England where some further development occurred. As the showers moved north they organised into bands of rain orientated west-south-west to east-north-east before gradually dying out in the evening. Later in the period, further showery rain moved southeast from the far northwest corner of the domain.

4.1.1 Subjective assessment

Figure 3 shows the 6-hour accumulations from the 5 model configurations and from radar for the period 10 to 16 UTC. The forecasts started at 09 UTC. Over this period, the 12-km forecast (b) did not manage to produce the heavy showers over southeast England. The 4 and 1-km runs did, but the forecasts with no high-resolution data assimilation (c) and (d) produced too much rain. The 4-km run with data assimilation (e) and the 1-km run with added increments (f) both produced reasonable accumulations but the spatial accuracy does not appear to be as good as the no-assimilation runs.

It must be emphasised that the objective scores presented later are taken from all four cycles and this is only showing one of them. However, a scrutiny of all the forecasts (not shown) revealed that the characteristics just described seem to be evident in all four of the verification periods and it is expected that they should be seen in the objective scores.

The main conclusions from the subjective assessment are:

1. The 12-km gridlength forecasts with data assimilation failed to produce the heavy showers that were observed over southeast England.

2. The 4 and 1-km gridlength forecasts with no high-resolution data assimilation ('spin up' runs) produced too much rain, but were the most accurate forecasts in terms of spatial distribution of the higher accumulations.

3. The 4 and 1-km gridlength forecasts with data assimilation produced the most accurate precipitation accumulations but were not as spatially accurate as the 4 and 1-km spin up runs.

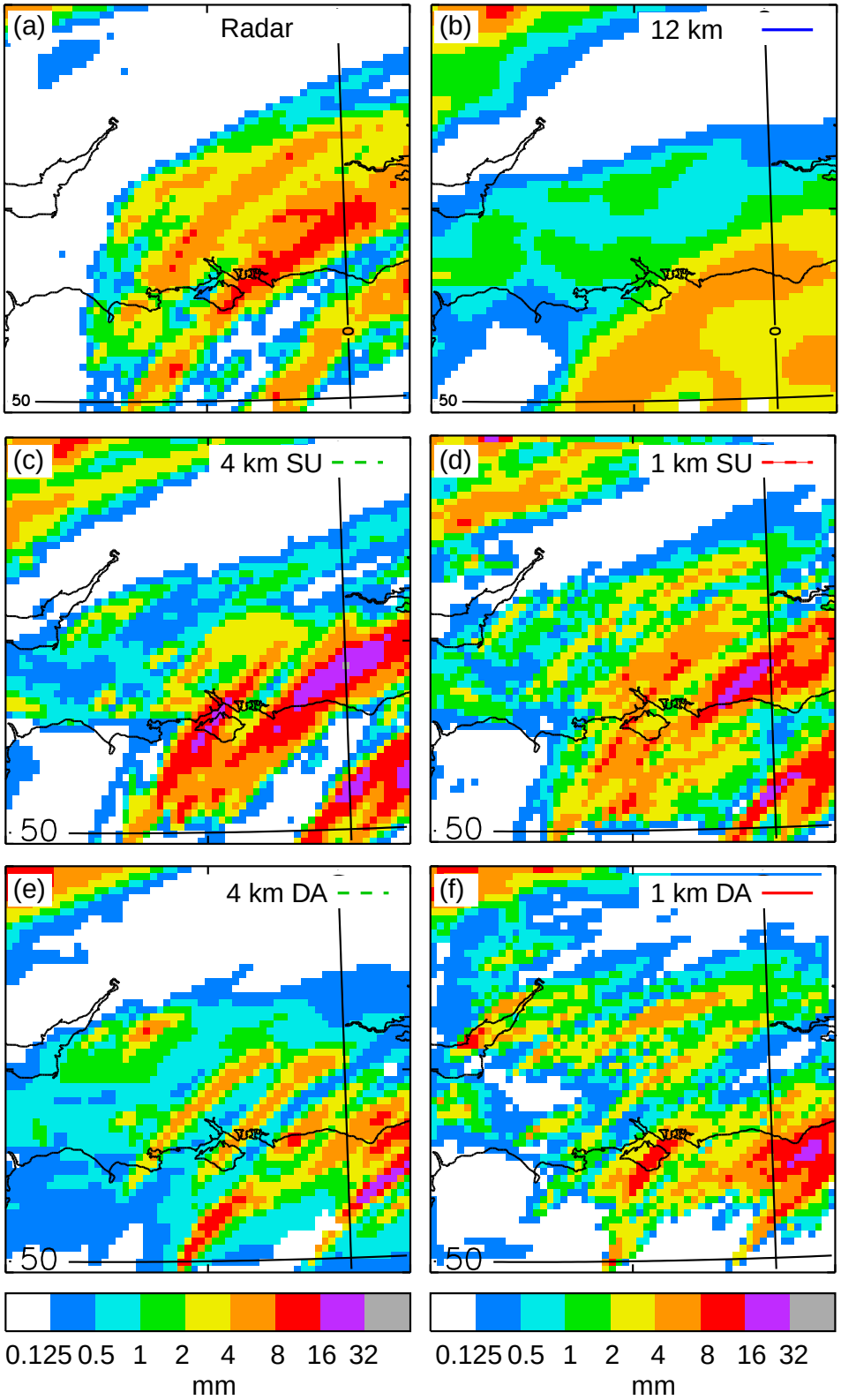


Figure 3. Rainfall accumulations over the period 10 to 16 UTC 28th August 2003 from (a) Nimrod radar composite. Forecasts starting at 09 UTC (a) 12-km gridlength with data assimilation, (b) 4-km gridlength 'spin up' starting from 12-km fields, (d) 1-km gridlength 'spin up' starting from 12-km fields, (e) 4-km gridlength with data assimilation and (f) 1-km gridlength with data assimilation.

4.2 13th May 2003 06 to 22 UTC

Early on, scattered showers were confined to the northwest of the domain. By mid-morning further scattered showers had developed widely and become heavy. During the afternoon the showers became more organised and thunderstorms were detected. From around 15 UTC a more intense line of thunderstorms generated over central southern England and into the southeast with locally very heavy rain and hail. This line persisted for a couple of hours, moving slowly southeastwards. By 19 UTC most of the showers had died away.

4.2.1 Subjective assessment

Figure 4 shows the 6-hour accumulations from the 5 model configurations and radar for the period 10 to 16 UTC. The forecasts started at 09 UTC. Over this period, the 12-km forecast (b) did not manage to produce the larger accumulations over southeast England. The showers were too light, died out too early and did not extend over the sea, but the general region of showery activity was reasonable. The 4 and 1-km runs with no high-resolution data assimilation (c) and (d) did produce more realistic rainfall totals because they were able to organise the showers in the afternoon. The 4-km run with data assimilation (e) was similar to the 4-km spin-up runs, but with slightly higher accumulations. The 1-km run with added increments (f) produced the best forecast of the higher accumulations, both spatially and in amount.

Again, a scrutiny of all the forecasts (not shown) revealed that the characteristics just described seem to be evident in all four of the verification periods and it is expected that they should be seen in the objective scores.

The main conclusions from the subjective assessment are:

1. The 12-km gridlength forecasts with data assimilation failed to produce the heavy organised showers that were observed in the afternoon over southeast England, but did get the general area of showery activity correct (except over the sea).

2. The 4 and 1-km gridlength forecasts with no high-resolution data assimilation ('spin up' runs) and the 4-km data assimilation runs produced better forecasts of the higher accumulations because they were able to maintain and organise showers into the afternoon.

3. The 1-km gridlength forecast with data assimilation produced the most accurate forecasts. In particular, it made the best attempt at simulating the band of heavy showers over southern/ southeast England in the afternoon.

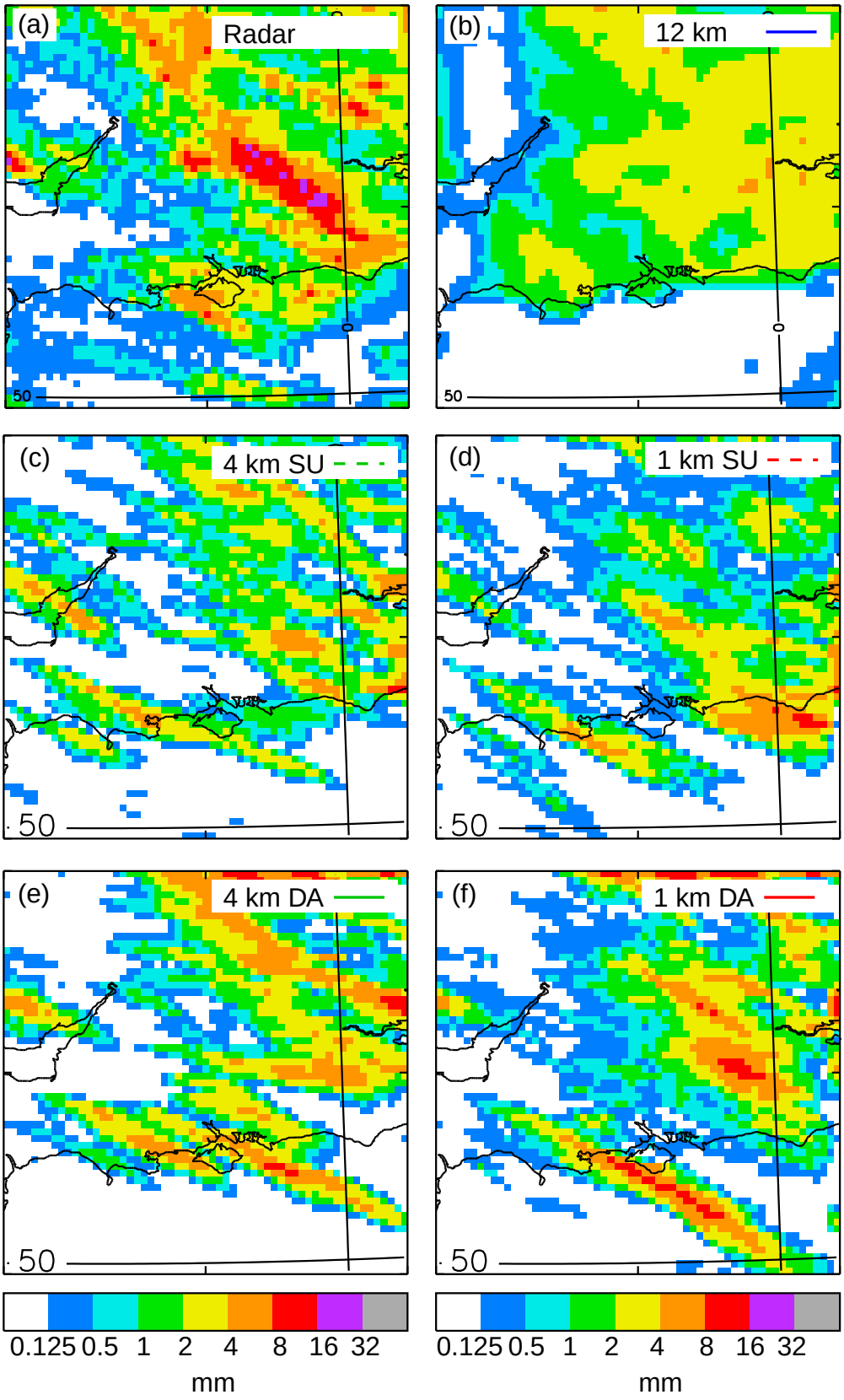


Figure 4. Rainfall accumulations over the period 10 to 16 UTC 13th May 2003 from (a) Nimrod radar composite. Forecasts starting at 09 UTC (a) 12-km gridlength with data assimilation, (b) 4-km gridlength 'spin up' starting from 12-km fields, (d) 1-km gridlength 'spin up' starting from 12-km fields, (e) 4-km gridlength with data assimilation and (f) 1-km gridlength with data assimilation.

4.3 25th May 2003 06 to 22 UTC

Scattered showers developed in a northwesterly flow in the morning. They developed further in the early afternoon before dying out from mid-afternoon onwards. Some organisation did occur, and some heavier bands of showers were observed over the far southeast of England for a while.

4.3.1 Subjective assessment

Figure 5 shows the 6-hour accumulations from the 5 model configurations and radar for the period 10 to 16 UTC. The forecasts started at 09 UTC. Over this period, the 12-km forecast (b) produced a good forecast. It generated showers over the correct area. Its biggest shortfall was that it did not manage to produce the higher accumulations over the far south of England because the showers died out too early. The 4 and 1-km runs with no high-resolution data assimilation (c) and (d) produced rainfall totals that were far too high, especially the 4-km runs, which developed unrealistic looking showers. The 4 and 1-km runs with data assimilation showed similar behaviour to the equivalent 'spin up' runs, but with more realistic totals.

Again, as for the previous two cases, a scrutiny of all the forecasts (not shown) revealed that the characteristics just described seem to be evident in all four of the verification periods and it is expected that they should be seen in the objective scores.

The main conclusions from the subjective assessment are:

1. The 12-km gridlength forecasts with data assimilation did a good job of predicting the overall distribution of the rain. They did not get the highest accumulations and did not maintain shower activity for long enough.

2. The 4 and 1-km gridlength forecasts with no high-resolution data assimilation ('spin up' runs) produced too much rain. The 4-km runs developed unrealistic looking showers. The spatial distribution of the higher precipitation amounts in the 1-km runs appeared to be marginally the best of all the forecasts

3. The 4 and 1-km gridlength forecasts with data assimilation produced less over-production of rain than the equivalent 'spin up' runs, but were otherwise similar.

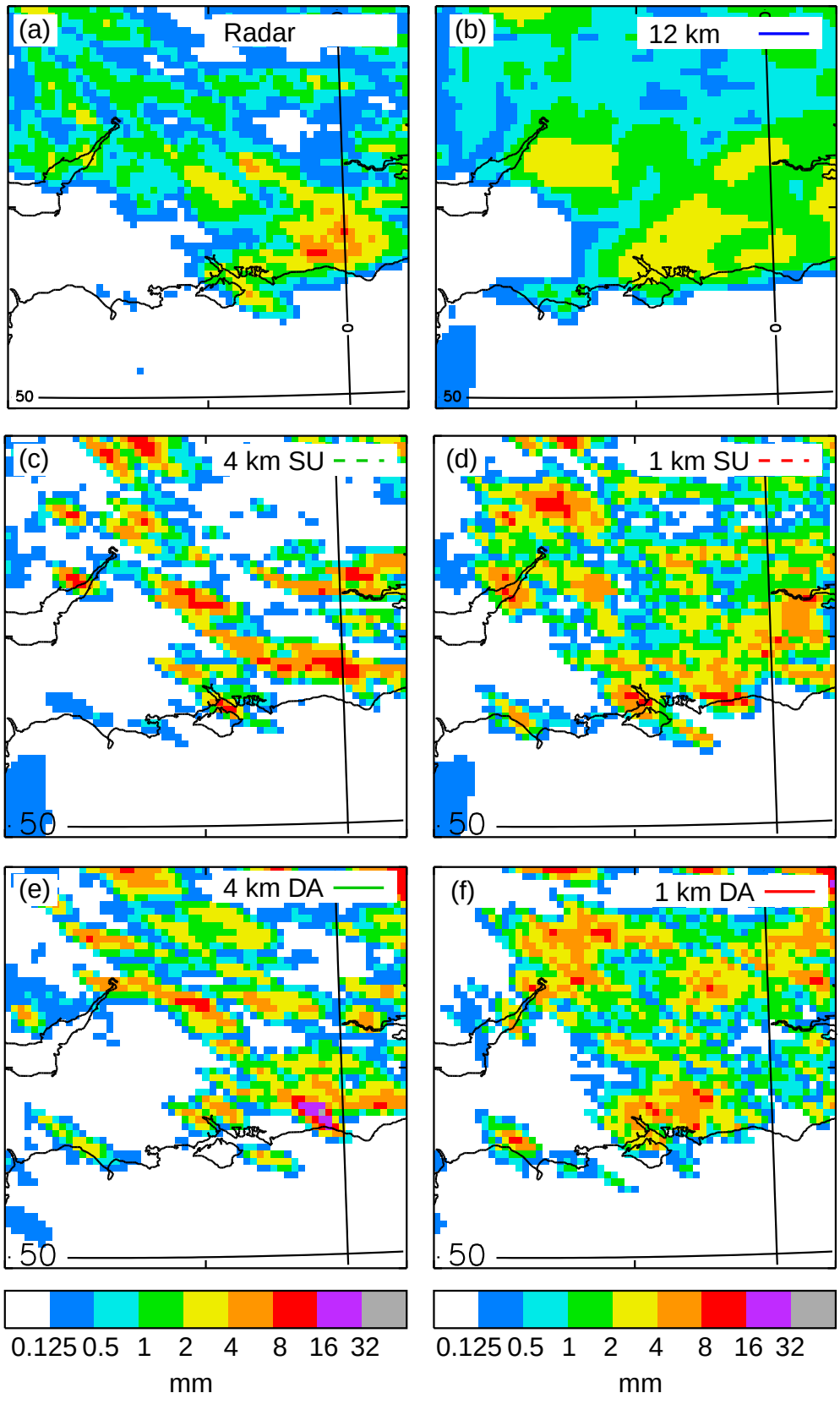


Figure 5. Rainfall accumulations over the period 10 to 16 UTC 25th May 2003 from (a) Nimrod radar composite. Forecasts starting at 09 UTC (a) 12-km gridlength with data assimilation, (b) 4-km gridlength 'spin up' starting from 12-km fields, (d) 1-km gridlength 'spin up' starting from 12-km fields, (e) 4-km gridlength with data assimilation and (f) 1-km gridlength with data assimilation.

4.4 1st July 200306 to 22 UTC

During the morning bands of southwest-northeast orientated showers developed over southeast and centra1 southern England. They reached their peak intensity around the middle of the day and then gradually weakened and moved away southeast during the afternoon. More showers moved into central southern England during the evening and intensified for a while in a northwest-southeast orientated line before moving southeast and decaying.

4.4.1 Subjective assessment

Figure 5 shows the 6-hour accumulations from the 5 model configurations and radar for the period 10 to 16 UTC. The forecasts started at 09 UTC. Over this period, the 12-km run (b) produced a generally good forecast. It generated bands of showers over the correct area, although a little to extensively, and also got the correct accumulations. It did not reproduce the later showers very well. The 4 and 1-km runs with no high-resolution data assimilation (c) and (d) produced rainfall totals that were far too high. They developed showers that were far too intense. However, the 1-km runs did generate the bands of showers in a more accurate location than the other forecasts. The 4 and 1-km run with data assimilation showed similar behaviour to the equivalent 'spin up' runs, but with slightly more realistic totals – but neither was as accurate spatially as the 1-km 'spin up'.

Again, as for the previous cases, a scrutiny of all the forecasts (not shown) revealed that the characteristics just described seem to be evident in all four of the verification periods and it is expected that they should be seen in the objective scores.

The main conclusions from the subjective assessment are:

1. The 12-km gridlength forecasts with data assimilation did a good job of predicting the rainfall accumulations that were observed. They generated showers over too wide an area, but were still good forecasts.

2. The 4 and 1-km gridlength forecasts with no high-resolution data assimilation ('spin up' runs) produced far too much rain. The spatial distribution of the higher precipitation amounts in the 1-km runs was the best out of all the forecasts.

3. The 4 and 1-km gridlength forecasts with data assimilation produced slightly less over-production of rain than the equivalent 'spin up' runs, but neither were as accurate spatially as the 1-km 'spin up' runs.

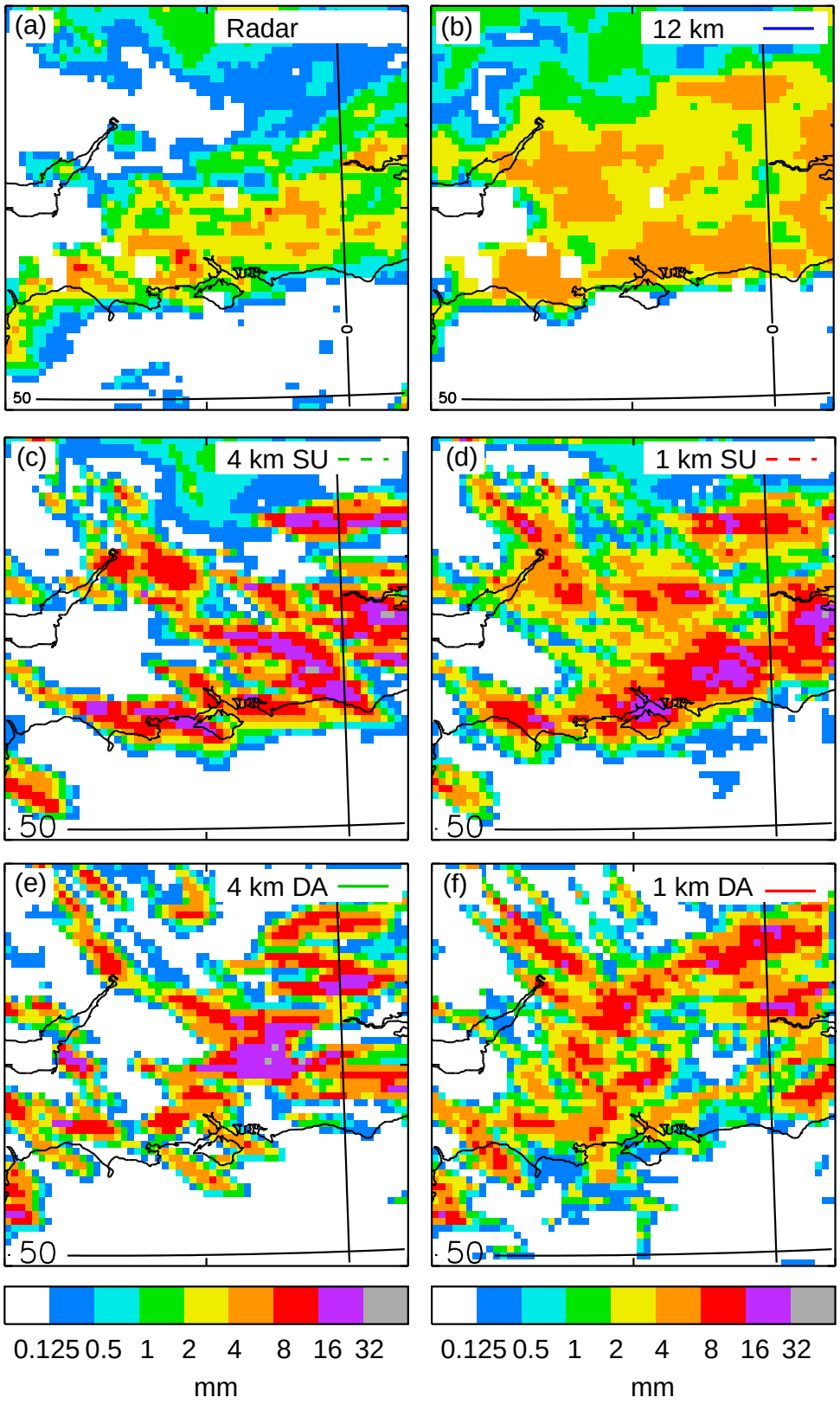


Figure 6. Rainfall accumulations over the period 10 to 16 UTC 1[tst] July 2003 from (a) Nimrod radar composite. Forecasts starting at 09 UTC (a) 12-km gridlength with data assimilation, (b) 4-km gridlength 'spin up' starting from 12-km fields, (d) 1-km gridlength 'spin up' starting from 12-km fields, (e) 4-km gridlength with data assimilation and (f) 1-km gridlength with data assimilation.

5 The objective verification approach

The verification was performed on a grid that covered the same area as the 1 km domain, but with a 5 km spacing between grid points. All the forecast and radar accumulations were either averaged or interpolated to this new verification grid before any scores were computed.

The decision to compare forecasts with radar rather than rain gauges was made because of the extra spatial information that radar provides. Although the error in radar-estimated rainfall at a single point may be larger than a gauge measurement, the representivity error over the features or regions of interest is less. For the purposes of this investigation, the radar data are taken to be 'truth' and free of error. The radar measurements had previously been processed through the Nimrod system to remove as much spurious rainfall as possible.

The aim of the verification was to:

1. Examine the skill of each of the forecast configurations to see the impact of data assimilation and resolution.
2. Examine the skill of each of the forecast configurations over different spatial scales.

In order to determine the skill over different spatial scales, the fractions of grid points within different sized squares that exceeded particular accumulation thresholds were computed.
The approach is as follows: Take a precipitation accumulation forecast on the verification grid (60x60 km pixels of size 5x5 km). Every pixel that exceeds the required threshold (say 4 mm) is classified as an event, all other pixels are non-events. For a particular pixel, compute the fraction of events that occur within a 3x3 pixel square centred on that chosen pixel. Do the same for every pixel. Then repeat for 5x5 squares, 7x7 squares, 9x9, 11x11 .....until the square size covers the whole domain. For each size of square used, a different field of fractions with values between 0 and 1 will have been generated. These fractions can then be compared (using suitable verification scores) with fractions generated from radar accumulations that have been computed in exactly the same way. The difference in skill over different spatial scales can be found by comparing verification scores obtained for different sized squares. A more detailed description of the approach is provided in the stage4 report.

Fractions were computed for 6-hour and 1-hour accumulations and compared with radar fractions using two scores – the Fractions Skill Score (FSS) and the log Odds Ratio (θ). Forecast fractions were also compared with radar binary values (1 where the threshold is exceeded, 0 if it is not) using the Brier skill score (BSS). Graphs of these scores will be presented. The scores and their characteristics are briefly described below. More detailed information is provided in Appendix B and in the stage4 report.

Another measure that will be presented is the mean accumulation over the whole domain. This shows the bias in forecast rainfall accumulations and whether a particular forecast configuration is over or under forecasting rainfall.

In section 6, scores will be shown for each of the events separately. They are made up of statistics from each of the 4 forecasts combined. In section 7, scores from all of the events combined will be shown.

5.1 Verification scores

5.1.1 Fractions skill score for comparing forecast fractions against radar fractions.

The purpose of this score is to examine the skill in forecasting rainfall areas that exceed a particular threshold. Forecast and radar fractions were compared over different spatial scales (different sized squares). Equations are given in Appendix B.
A perfect score has a value of 1.0 (observed and forecast fractions are in exact agreement). A forecast with no skill has a score of 0.0 (no fractions >0.0 overlap) – i.e. there is no spatial correlation between the forecast and observed fields. If no accumulations exceeding the threshold are forecast when there were some, or if accumulations exceeding the threshold were predicted when none occurred, the score will be 0.0. For any given accumulation threshold, the forecast skill should improve with spatial scale, but be limited to a value that is determined by the ratio of number of forecast pixels to observed pixels over the whole domain.

5.1.2 The Brier skill score

This is the standard Brier skill score (see Appendix B and the stage 4 report). The purpose here is to determine the spatial skill of rainfall predictions. Equations are given in Appendix B. The fractions (they can be thought of as probabilities) are compared with a binary 1.0 at grid squares when the threshold is exceeded and 0.0 where it is not. The 'climatology' used here is the score that would be obtained if the forecast produced random precipitation with the same number of pixels exceeding the threshold as that observed. It is only meaningful, when using this random climatology with the fractions approach, to use thresholds which compare say the top 5% or 10% (95^{th} or 90^{th} percentile) of all grid points rather than have an absolute value (e.g. 4 mm). In this way, the locations of the highest forecast accumulations can be compared with the highest observed accumulations over different spatial scales (different square sizes). When plotted against spatial scale, the curves will eventually asymptote to a value of 0.5. It is the spatial scale at which the curve has a value >0.5 (if at all) and the spatial scale at which the curve has a peak (if there is one) that are of most interest.

5.1.3 The log Odds Ratio (comparing forecast and radar fractions)

The purpose is to examine the skill in rainfall accumulation forecasts in terms of the ratio of the odds of forecasting an accumulation exceeding the threshold being correct to the odds of forecasting an accumulation exceeding the threshold being wrong (see Stephenson 2000). Forecast and radar fractions were compared over different spatial scales. The score is described more fully in the stage 4 report and in Appendix B. A forecast with no skill has a negative score. A score of 1 means odds of 10:1 for getting the forecast correct, a score of 2 means odds of 100:1 for getting a forecast correct. The odds ratio tends to give better scores for rarer events (higher thresholds).

6 Verification scores for each event

Curves for the 5 model configurations are always displayed by the line colour and style shown in the key below.

1 km gridlength	spin up from 12 km
4 km gridlength	spin up from 12 km
1 km gridlength	added increments from 4 km VAR assimilation
4 km gridlength	4 km VAR assimilation + 4 km MOPS
12 km gridlength	12 km VAR assimilation + 12 km MOPS
radar	

6.1 Scores for 6-hour accumulations

6.1.1 Fractions skill scores

Results from the low accumulation threshold of 0.001 mm (Figure 7)

1. All the forecasts have high scores and perform almost the same. This is not surprising; the forecasts of getting at least some rain over a 6 period appeared to be very good for all the events.
2. The scores all increase the most at small spatial scales (small squares, up to ~25 km) and then level off, reaching their asymptotic value by ~100 km. We should expect that the poorest forecast is at the gridscale.

Results from the accumulation threshold of 4 mm (Figure 8)

1. There is much more variability in skill from event to event and from forecast to forecast than there was for the very low threshold. This is partly due to the nature of the fractions skill score.
2. The scores are lower than they were for the very low threshold because a higher threshold is more difficult to predict.
3. The skill at the gridscale (far left of graph) is generally poor. Skill improves rapidly with spatial scale for most forecasts, before gradually tending to an asymptotic value at the larger spatial scales.
4. The 12-km gridlength forecasts (blue line) are very poor in events (B) and (C), but are the best in event (D). This agrees with the subjective assessment earlier. The reason for the poor forecasts is that insufficient precipitation was generated.

5. The 1 and 4 km forecasts with data assimilation appear to have similar behaviour to each other as do the 1 and 4 km forecasts that spin up from the 12 km T+1.
6. There is no obvious overall best set of forecasts. However, the 1 km with added increments has the best skill in event (B) and the 1 and 4 km spin up set of forecasts are the most skilful in event (A). These two are the most important of the four events as they produced the highest accumulations locally and are more like the sort of meteorological situation a storm scale model will be designed for.

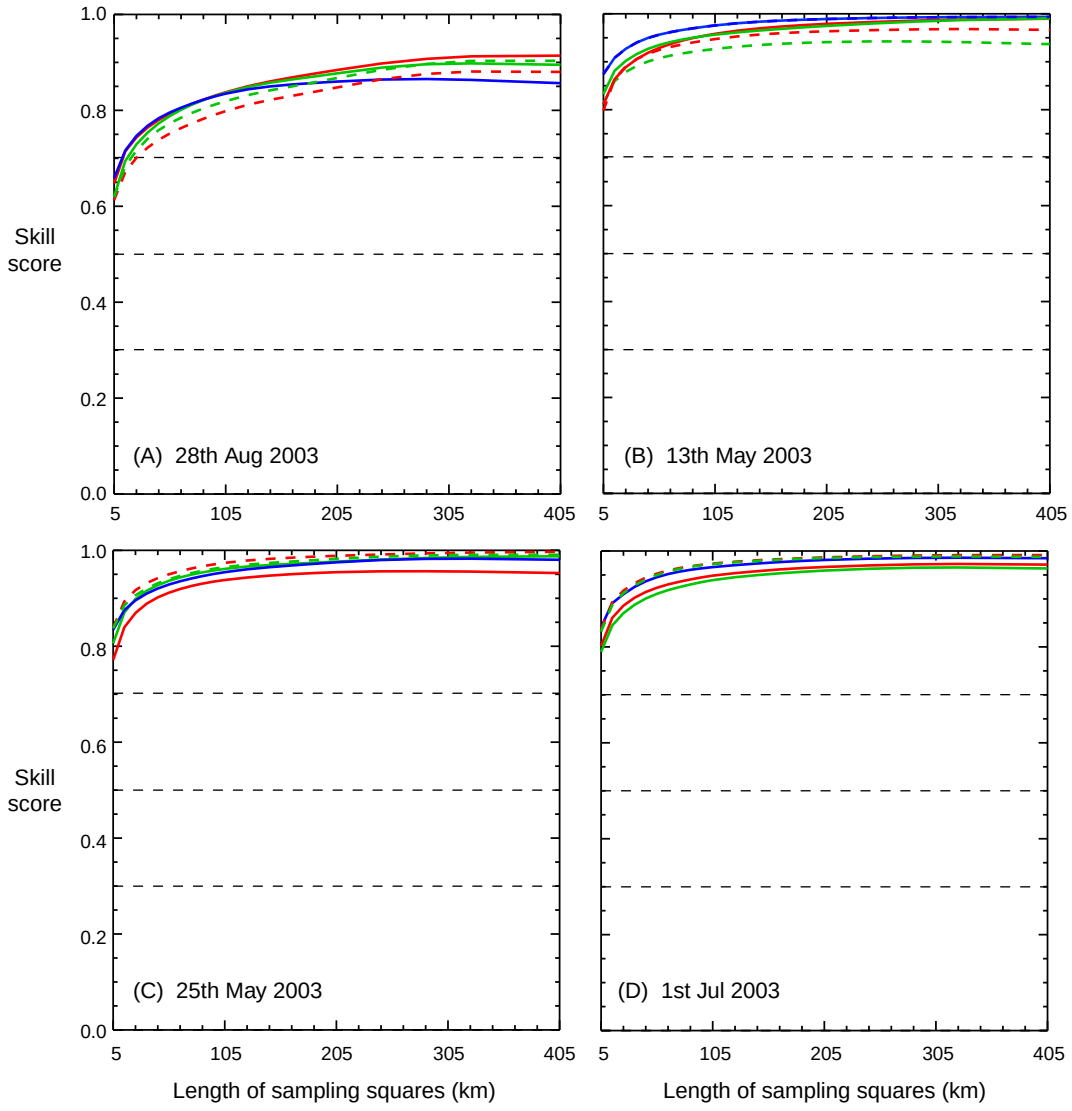


Figure 7. Skill scores for fractions of pixels with 6-hour accumulations exceeding 0.001 mm within different sized squares for each of the events. The bottom axis shows the length of the sides of the squares over which the forecast and radar fractions were computed. A score of 1.0 indicates a perfect forecast for a given square size; a score of 0.0 indicates no skill for a given square size. Further information on this skill score is given in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

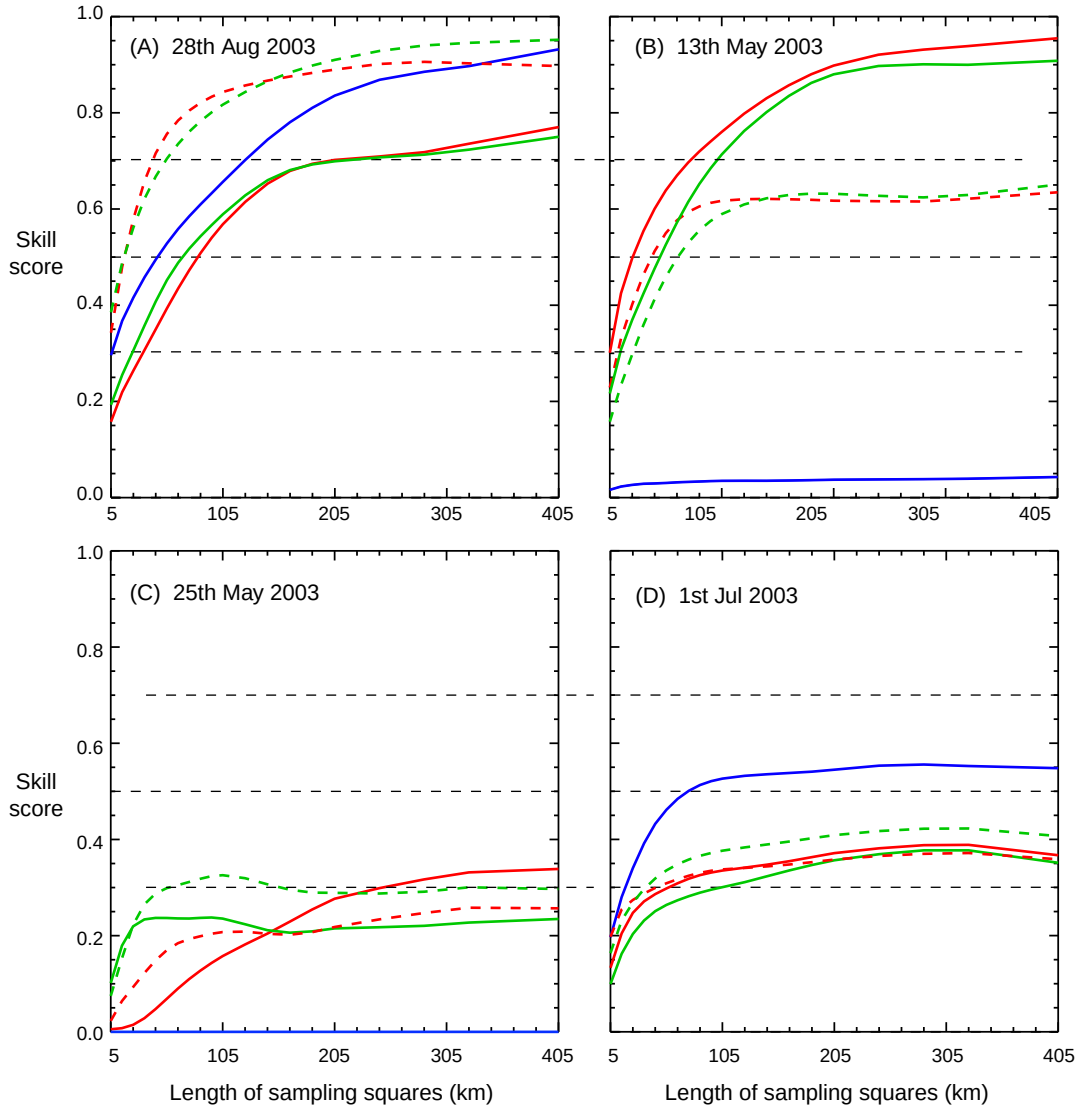


Figure 8. Skill scores for fractions of pixels with 6-hour accumulations exceeding 4.0 mm within different sized squares for each of the events. The bottom axis shows the length of the sides of the squares over which the forecast and radar fractions were computed. A score of 1.0 indicates a perfect forecast for a given square size; a score of 0.0 indicates no skill for a given square size. Further information on this skill score is given in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

Results from the accumulation threshold of 10% or 90th percentile value (Figure 9)

1. All the curves tend to a perfect forecast for the largest spatial scales. This is exactly what should happen when the forecast and observed frequencies are defined as the same (i.e. 0.1) when computing the score.
2. A trend can now be seen. The 1 and 4 km 'spin up' forecasts (dashed lines) are the best, or close to the best, in all the events. This is saying that the spatial distribution of the highest accumulations in the 'spin up' forecasts is better than in the rest – but it still may be that the bias is worse.

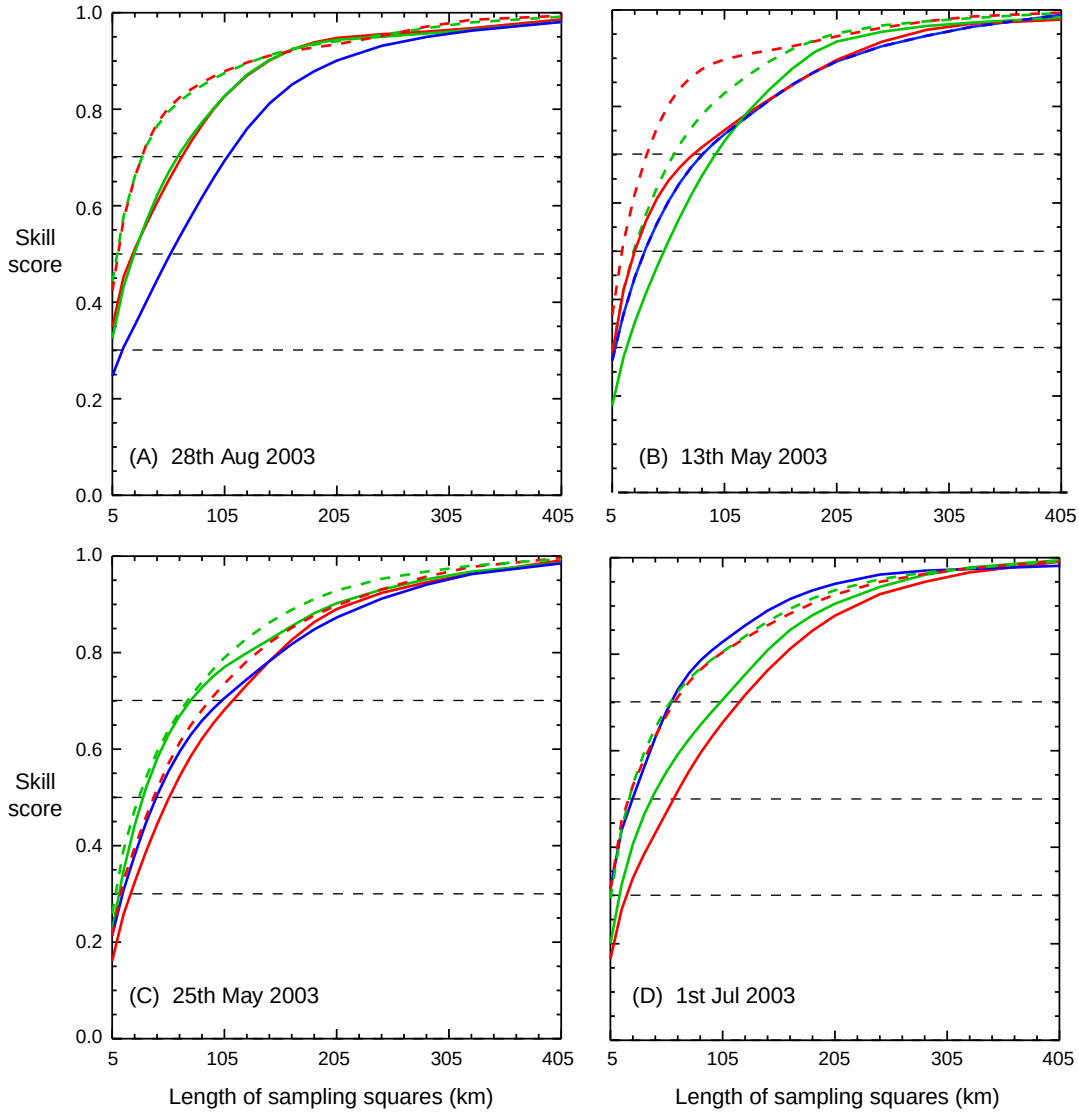


Figure 9. Skill scores for fractions of pixels with 6-hour accumulations exceeding a threshold of 10% within different sized squares for each of the events. The threshold of 10% is the lowest accumulation value of the highest 10% of pixels within the entire domain – the 90th percentile value. The bottom axis shows the length of the sides of the squares over which the forecast and radar fractions were computed. A score of 1.0 indicates a perfect forecast for a given square size; a score of 0.0 indicates no skill for a given square size. Further information on this skill score is given in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

6.1.2 Brier skill scores

The Brier skill scores (Figure 10) are displayed for a threshold of 10% (90th percentile).

Key results

1. They are comparable to the skill scores shown in Figure 9, except that this measure gives a different view of the variation of skill with spatial scale. The graph reinforces the previous conclusion that the 1 and 4 km 'spin up' forecasts are the best for this frequency threshold for all the events. Their curves (dashed) cross the 0.5 line at a smaller spatial scales (smaller squares) and have a higher peak than the others.

2. The 12-km forecasts come out worst overall, which backs up the previous subjective assessment that the 12-km forecasts produce smoother, more bland, accumulation fields than the rest and are therefore less likely to come out with the best spatial distribution of the highest 10% of accumulations

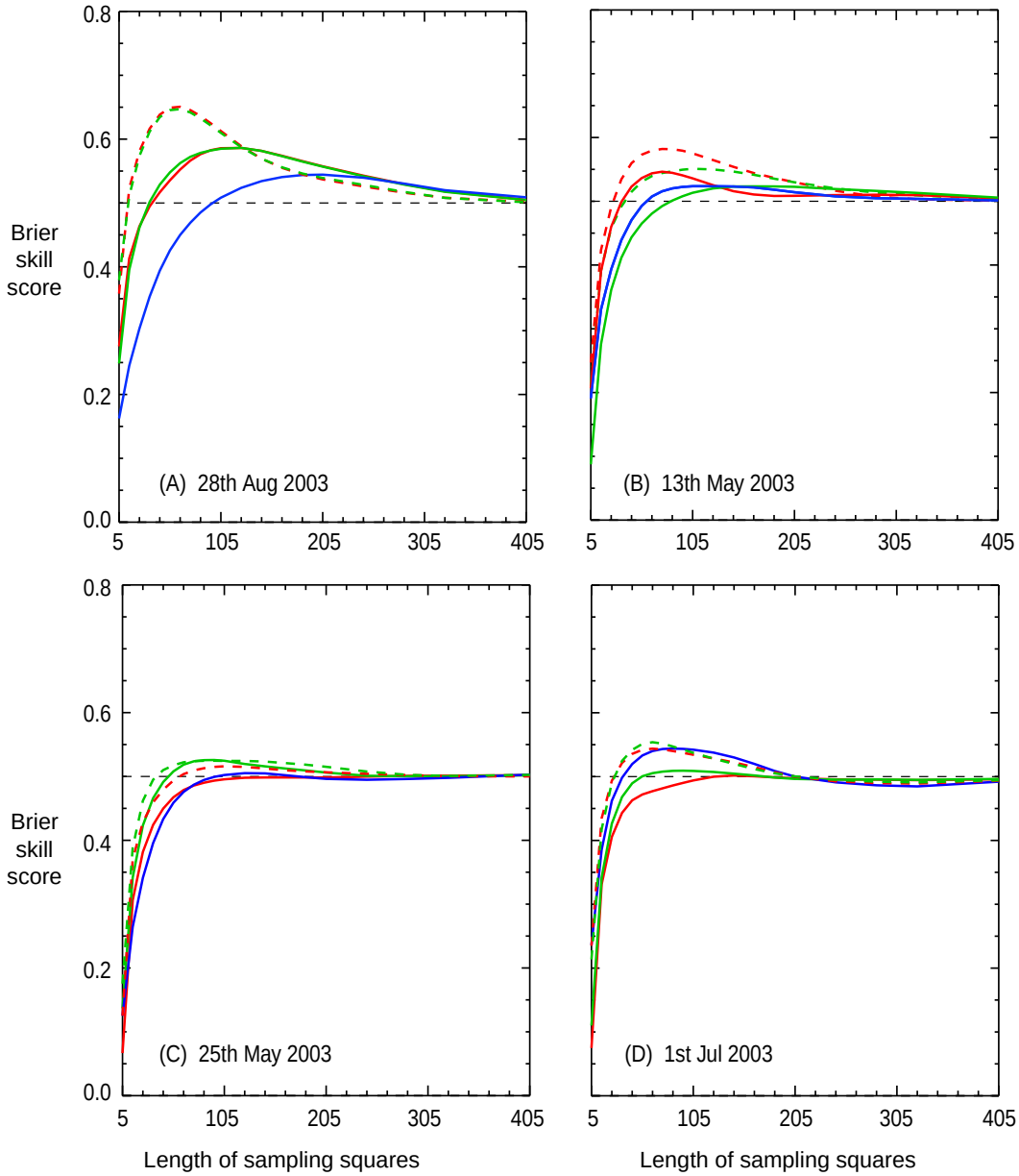


Figure 10. Brier skill scores for forecast fractions of pixels with 6-hour accumulations exceeding a threshold of 10% within different sized squares for each of the events, compared to radar values of 1.0 if the threshold is exceeded or 0.0 if it is not. The threshold of 10% is the lowest accumulation value of the highest 10% of pixels within the entire domain – the 90th percentile value. The bottom axis shows the length of the sides of the squares over which the forecast and radar fractions were computed. A score greater than 0.5 indicates skill better than that obtained from sampling over the whole domain. Further information on this Brier skill score is given in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

6.2 Scores for hourly accumulations

6.2.1 Mean accumulations

Figure 11 shows graphs of the mean hourly accumulations for each event using a square size of 55 x 55 km. The first points '1 to 2' are the hour from T+1 to T+2, '2 to 3' is the hour from T+2 to T+3 etc. The key results are:

1. The 1 and 4 km 'spin up' runs (dashed lines) always started with too little rain, but the amount of rain increased markedly through the course of the forecasts.
2. In general (apart from event (B)), the 1 and 4 km 'spin up' runs produced too much rain – even though the spatial distribution of the highest precipitation areas was the best in those forecasts. This agrees with the earlier subjective view.
3. The amount of rain produced by the 4 km forecasts with data assimilation (solid green lines) and the 1 km forecast with added increments (solid red lines) tended to decrease with forecast time (except event (C)).

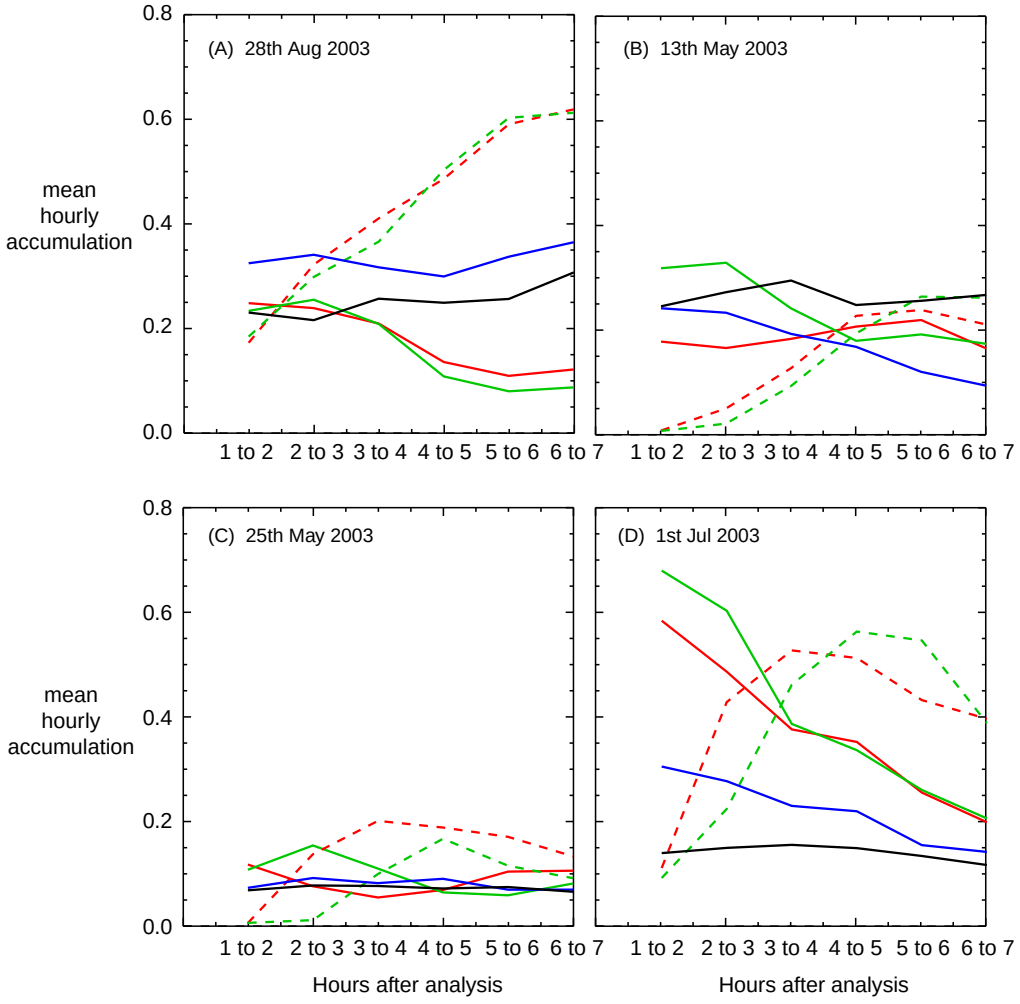


Figure 11. The mean hourly accumulations over the whole domain (in mm) for each event. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. The black line = radar, blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

4. The 12-km forecasts produced mean hourly rainfall amounts that were closest to radar throughout the forecasts (except event (B)) and they were closest to the radar in three of the four events in the first hour of the forecasts.

6.2.2 Skill scores for comparison of fractions – key results

Results from the low accumulation threshold of 0.001 mm (Figure 12)

1. In two of the events (B) and (C) the 1 and 4 km 'spin up' forecasts (dashed lines) started with poor scores in the first hour. This is because they were producing very little rain at the start of the forecasts.
2. For events (A) and (D), the scores from all the forecasts were comparable and stayed almost constant with time.
3. The 12 km forecasts were marginally better in three of the events

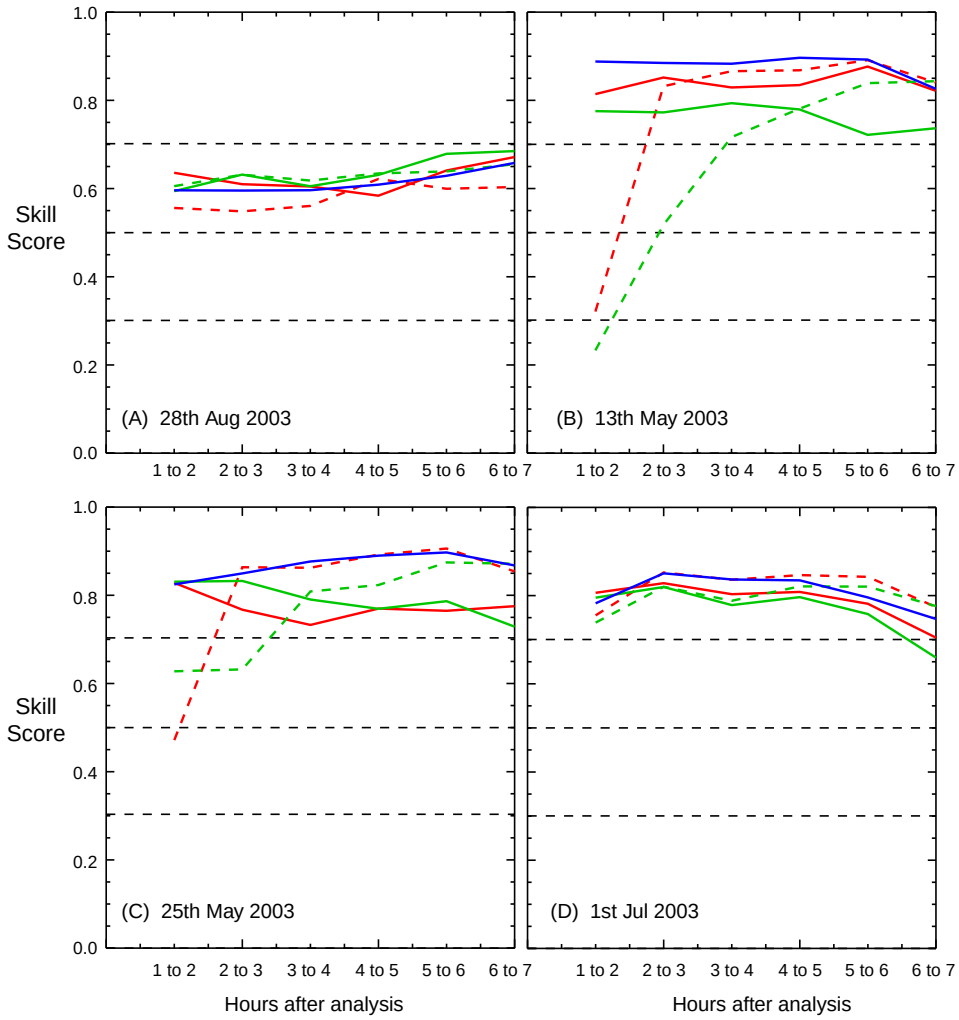


Figure 12. Skill scores for fractions of pixels with hourly accumulations exceeding 0.001 mm within squares of 55x55km for each of the events. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. A score of 1.0 indicates a perfect forecast for this square size; a score of 0.0 indicates no skill for this square size. Further information on this skill score is given in the stage 4 report (Roberts 2004). Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

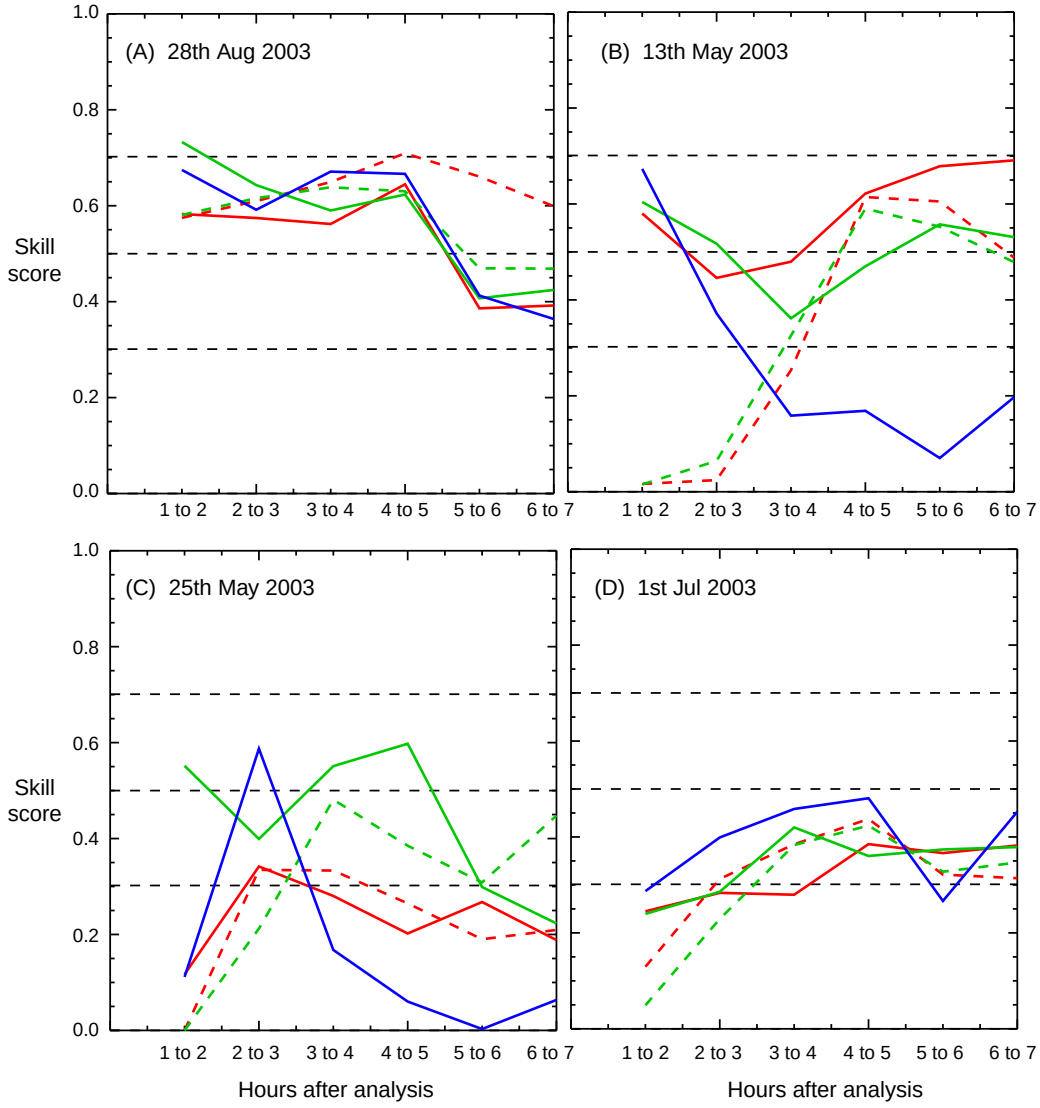


Figure 13. Skill scores for fractions of pixels with hourly accumulations exceeding 1 mm within squares of 55x55km for each of the events. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. A score of 1.0 indicates a perfect forecast for this square size; a score of 0.0 indicates no skill for this square size. Further information on this skill score is given in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

Results from the accumulation threshold of 1 mm (Figure 13)

1. The range of skill between forecasts was much greater than for the very low accumulation threshold.
2. The skill of the hourly accumulation forecasts was generally not a great as for the 6-hour accumulation forecasts. This is no surprise, as timing errors in the forecasts become more important.
3. Over all the events, the 1 and 4 km 'spin up' forecasts started off with the poorest scores. Again, this is because they were producing too little rain at the start, and it was even more of a problem with a higher threshold.
4. The 12-km forecasts lost skill with forecast time in three of the events. In (B) and (C) this loss of skill was very dramatic. It was also noted in the subjective assessment that the 12-km forecasts tend to lose their showers too soon in some of the forecasts.

5. The forecasts from the 1 km runs with added increments (solid red line) in (B) and the 1 km 'spin up' runs in (A) (dashed red line) were significantly more skilful than the others. This is in agreement with the skill of those forecasts over 6-hours.

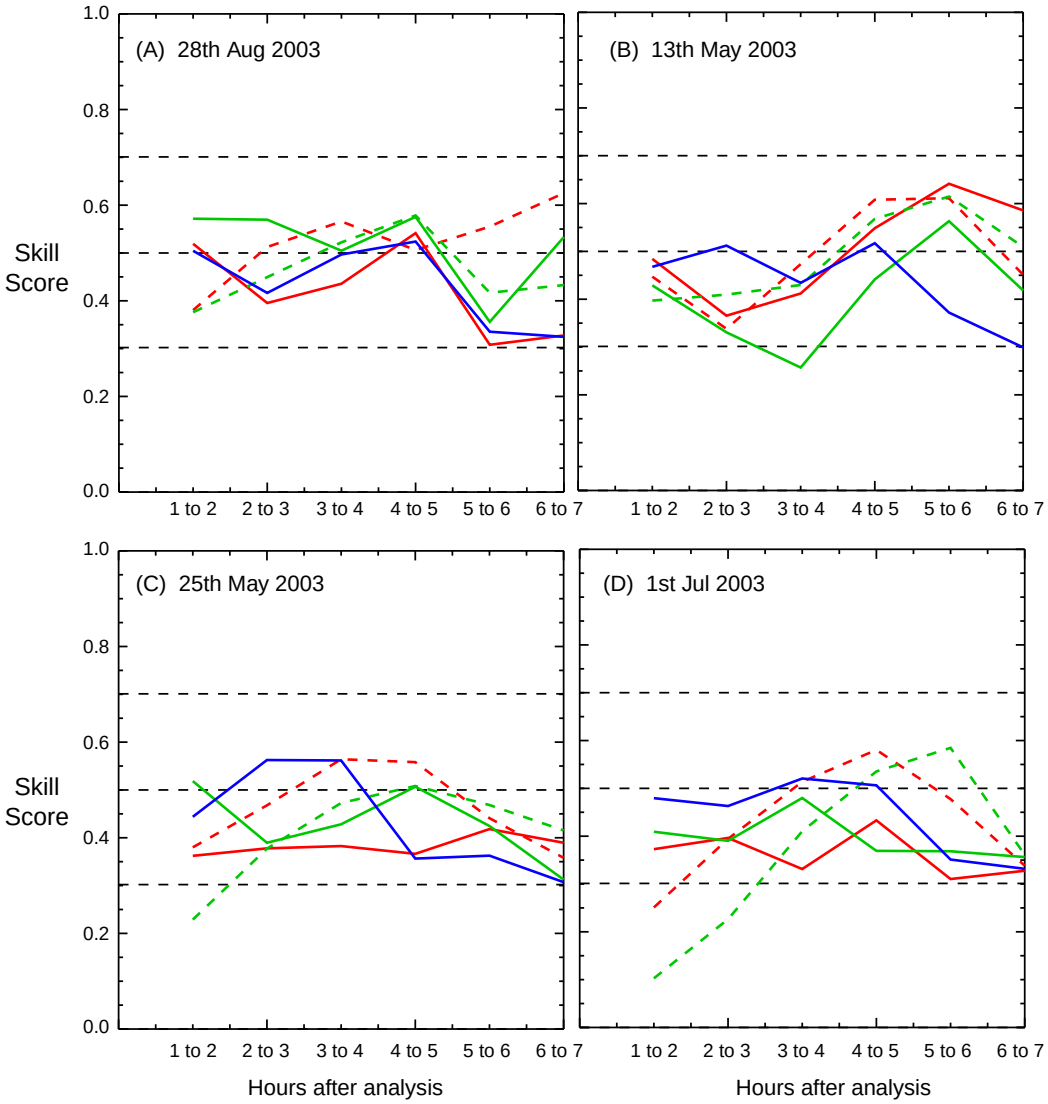


Figure 14. Skill scores for fractions of pixels with hourly accumulations a threshold of 10% within squares of 55x55km for each of the events. The threshold of 10% is the lowest value of the highest 10% of pixels within the entire domain – the 90th percentile value. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. A score of 1.0 indicates a perfect forecast for this square size; a score of 0.0 indicates no skill for this square size. Further information on this skill score is given in the stage 4 report (Roberts 2004). Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

Results from the accumulation threshold of 10% or 90th percentile value (Figure 14)

1. The 1 and 4 km 'spin up' forecasts had the poorest scores in the first hour, but were then generally the best (or close to the best) later in the forecast.
2. Apart from the initial increase in skill in the 1 and 4 km 'spin up' runs, the level of skill of the hourly accumulations remained roughly the same for all the forecasts over the 6 hours – though with some departures from that (e.g. the 1 km run with added increments improved in event (B) and the 12 km scores got worse for all the events.

7 Verification scores for all events

Curves for the 5 model configurations are always displayed by the line colour and style shown in the key below.

1 km gridlength	spin up from 12 km
4 km gridlength	spin up from 12 km
1 km gridlength	added increments from 4 km VAR assimilation
4 km gridlength	4 km VAR assimilation + 4 km MOPS
12 km gridlength	12 km VAR assimilation + 12 km MOPS
radar	

7.1 Scores for 6-hour accumulations

7.1.1 Skill scores for comparison of fractions and odds ratios

Figure 15

1. The skill scores (a) and the log odds ratios (b) for the low accumulation threshold of 0.001 mm are very similar for all the model configurations.
2. The skill scores for the accumulation threshold of 4 mm (c) show that over all the events, and smaller spatial scales, the 1 and 4 km 'spin up' forecasts (dashed lines) were the most skilful, the 1 km with added increments (red line) and 4 km with data assimilation were similar and the next most skilful, and the 12-km forecasts were the least skilful
3. The odds ratios for the accumulation threshold of 4 mm (d) do not show very much difference between the forecasts (unlike the skill scores mentioned above). They also show comparable skill to the forecasts of the low accumulation threshold, which may seem surprising when we know that higher thresholds are usually more difficult to forecast. However, it is a characteristic of the score that rarer events (i.e. higher thresholds) get better scores and this provides a compensating effect. It may also explain why the lines are close together: with the skill scores the 12 km forecasts are worse primarily because they produce too few pixels that exceed the threshold, the odds ratios will not penalise this so much. The log odds ratios have not been as useful as the fractions skill scores for discriminating between the forecasts.
4. The skill scores (e) (and to a lesser extent the log odds ratios (f)) show that for a threshold of the top 10% of grid squares (90^{th} percentile) the 1 and 4 km 'spin up' forecasts (dashed lines) were the most skilful over all the events. The others were comparable. The same result holds for thresholds of 25 and 5% (not shown). The 'spin up' forecasts produced the best spatial distribution of the highest accumulations.

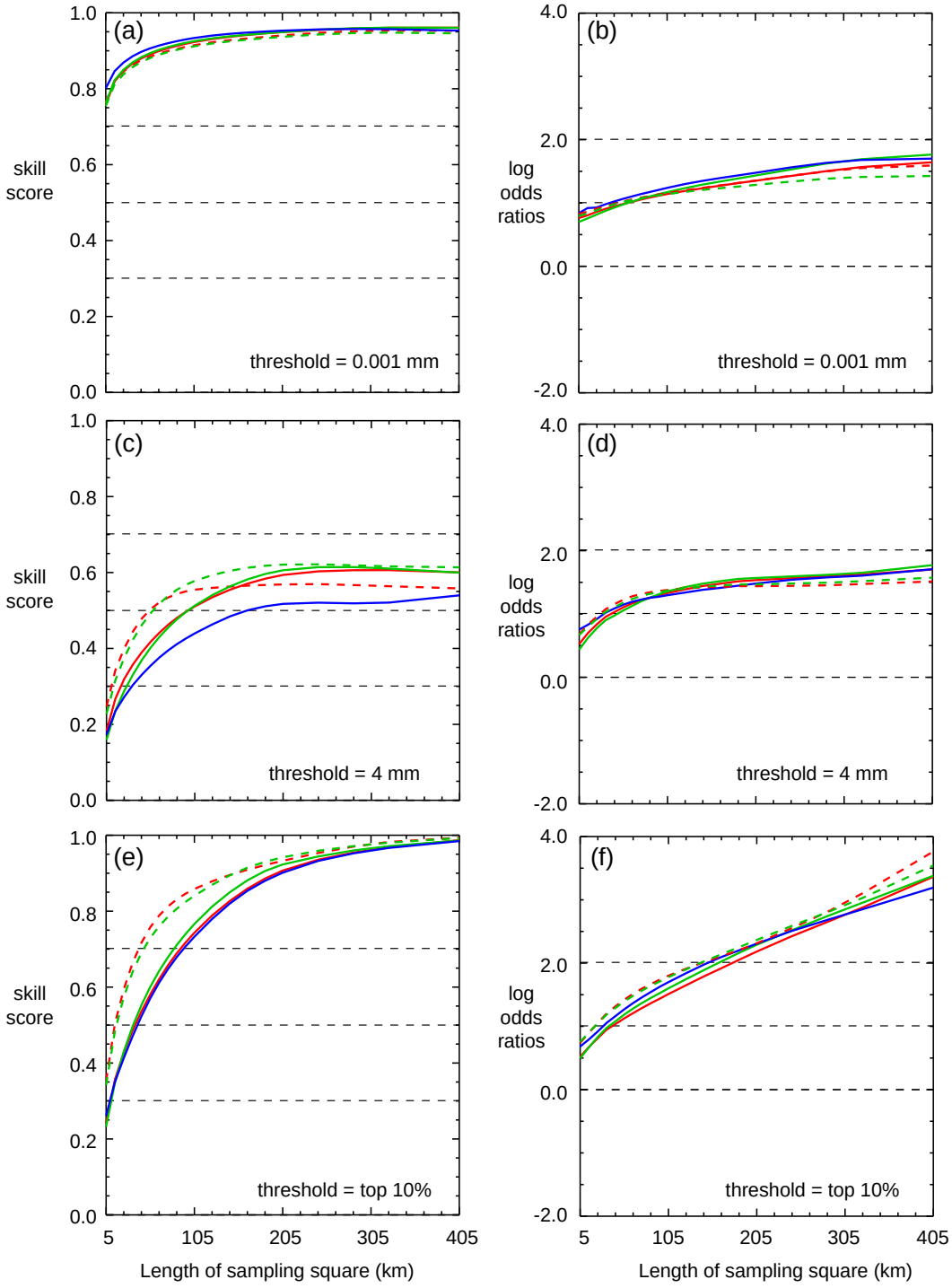


Figure 15. Scores for fractions of pixels with 6-hour accumulations exceeding thresholds within different sized squares for all four events. (a) Skill scores for a threshold of 0.001 mm; (b) log Odds ratios for a threshold of 0.001 mm; (c) Skill scores for a threshold of 4 mm; (d) log Odds Ratios for a threshold of 4 mm; (e) Skill scores for a threshold of 10%; (f) log Odds ratios for a threshold of 10%. The threshold of 10% is the lowest value of the highest 10% of pixels within the entire domain – the 90th percentile value. The bottom axis shows the length of the sides of the squares over which the forecast and radar fractions were computed. A skill score of 1.0 indicates a perfect forecast for a given square size; a score of 0.0 indicates no skill for a given square size. A log odds ratio of less than 0.0 indicates no skill, a value > ~1.0 is good. Further information on the skill score and odds ratios is provided in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

Figure 16 (8 mm threshold)

The fractions skill score and the log odds ratios tell the same story. The 12 km forecasts had very little skill at any spatial scale. This was because the 8 mm threshold was hardly ever exceeded (the reason for no odds ratio curve). The 4 and 1 km forecasts were comparable at the smaller scales, except for the 4 km with data assimilation, which was less skilful. The reason for the difference in the 4 km data-assimilation forecasts may have been due to the inclusion of MOPS at 4 km. Over larger scales the 1 km with added increments and the 4 km forecasts with data assimilation performed better than the 4 and 1 km 'spin up' runs. The spin-up runs were poorer over large scales because the fraction of points over the whole domain with an accumulation above 8 mm was too high. The limit in skill with spatial scale for all the 1 and 4 km forecasts was due to the over forecasting of high accumulations.

The main signal is that higher resolution has a lead to a significant improvement in the prediction of rainfall accumulations exceeding 8 mm over the four events. The 1 km gridlength forecasts with added increments produced the best scores.

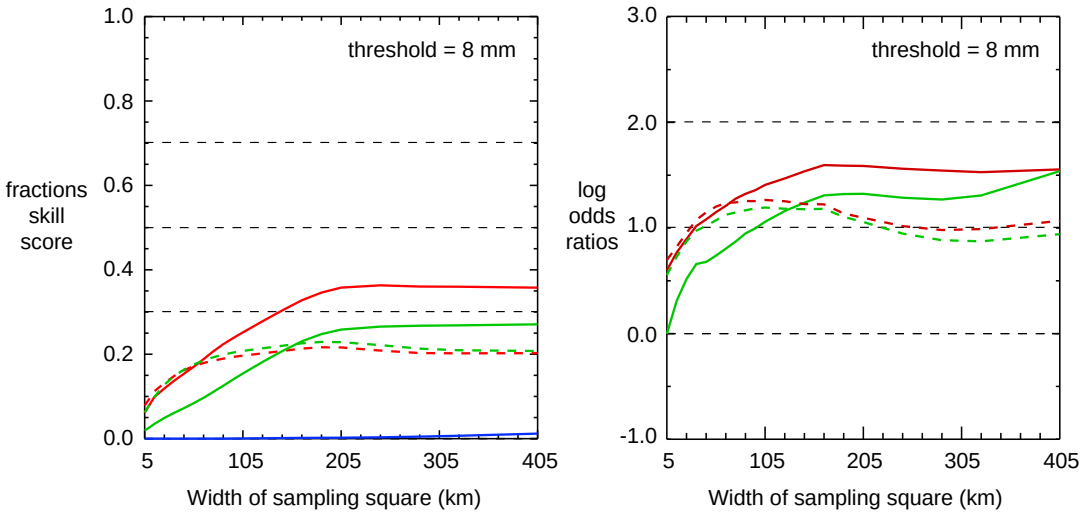


Figure 16. Fractions skill cores and log odds ratios for fractions of pixels with 6-hour accumulations exceeding a threshold of 8 mm for all four events.

7.1.2 Brier skill scores

The Brier skill scores are shown (Figure 17) for a threshold of the top 10% accumulations (90th percentile). The key results are:

1. As for the skill scores shown in Figure 15(e), the 1 and 4 km 'spin up' forecasts (dashed lines) were the most skilful overall. The 1-km with add increments (red line) and the 4-km with data assimilation (green line) were the next best. The 12-km forecasts were the worst.
2. A skill better than 0.5 (obtained from taking fractions over the whole domain) was reached for squares of side :
    (a) 20 km for the 1 km 'spin up' forecasts
    (b) 23 km for the 4 km 'spin up' forecasts
    (c) 53 km for the 1 km forecasts with added increments
    (d) 53 km for the 4 km forecasts with data assimilation
    (e) 70 km for the 12 km forecasts
3. The peak skill was found for squares of side :
    (a) ~65 km for the 1 km 'spin up' forecasts
    (b) ~65 km for the 4km 'spin up' forecasts
    (f) ~110 km for the 1 km forecasts with added increments
    (g) ~110 km for the 4 km forecasts with data assimilation
    (c) ~135 km for the 12 km forecasts

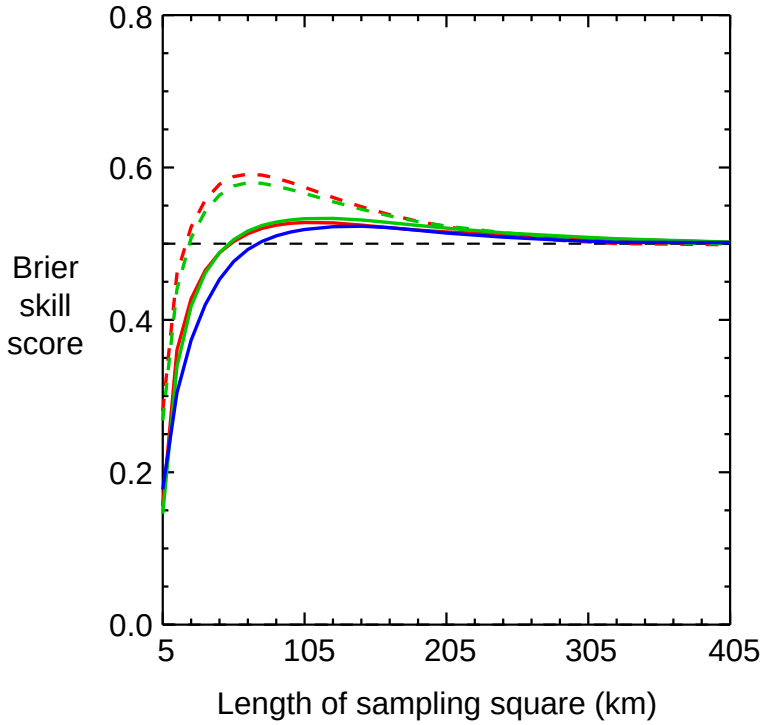


Figure 17. Brier skill scores for forecast fractions of pixels with 6-hour accumulations exceeding a threshold of 10% within different sized squares for all four events, compared to radar values of 1.0 if the threshold is exceeded or 0.0 if it is not. The threshold of 10% is the lowest value of the highest 10% of pixels within the entire domain – the 90th percentile value. The bottom axis shows the length of the sides of the squares over which the forecast and radar fractions were computed. A score greater than 0.5 indicates skill better than that obtained from sampling over the whole domain. Further information on this Brier skill score is given in Appendix B and in the stage 4 report. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

7.2 Scores for hourly accumulations

7.2.1 Skill scores for comparison of fractions and odds ratios

The key results from the graphs in Figure 18 are:

1. There is good agreement between the skill scores and the odds ratios in terms of the trends over the period for all three thresholds. The major difference comes with the scores for the first hour with the 4 mm threshold (c) and (d). The odds ratios did not have low values for the 1 and 4 km 'spin up' forecasts. This is again because the odds ratios can tend to favour the rarer events and give misleading results in some situations. The difference between the values in (b) and (d) show clearly how the odds ratios become generally higher for the rarer event (higher threshold).

2. For the very low accumulation threshold of 0.001 mm, the 12-km forecasts were the most skilful.

3. For the other thresholds the 12-km forecasts become the least skilful by the end of the forecast period.

4. With the threshold of the top 10%, the 1 and 4 km 'spin up' forecasts (dashed lines) have the least skill at the start of the forecasts, but become the best forecasts from the middle of the period onwards.

5. The 12-km and 4 km with data assimilation have the best scores over the first hour for the higher thresholds.

6. There seems to be a peak in skill at 4-5 hours in the 1 and 4 km forecasts for the higher thresholds. This may be a real signal, but could also be just an artefact from sampling a small number of forecasts.

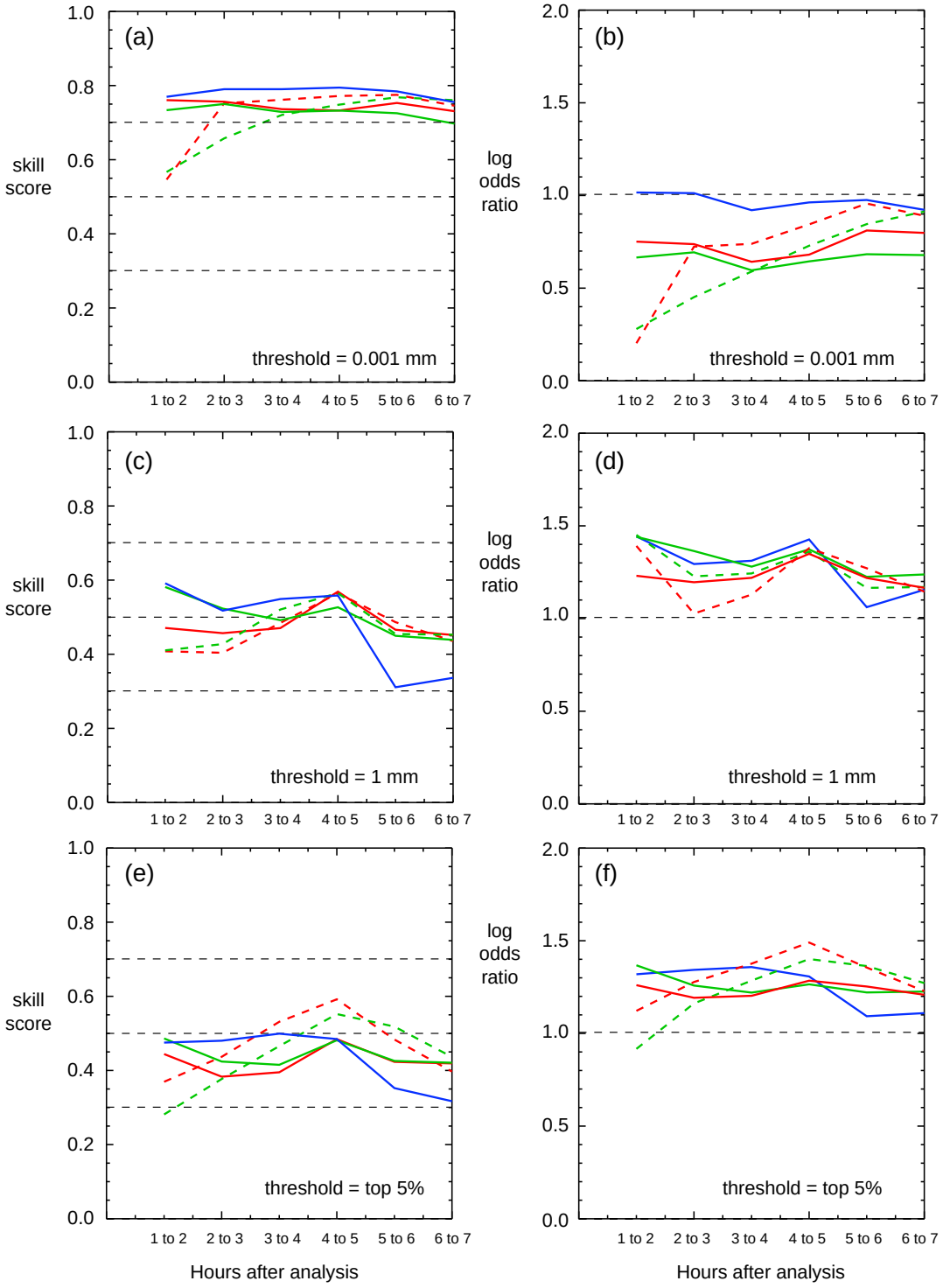


Figure 18. Scores for fractions of pixels with hourly accumulations exceeding thresholds within squares of size 55x55km for all four events. (a) Skill scores for a threshold of 0.001 mm; (b) log Odds ratios for a threshold of 0.001 mm; (c) Skill scores for a threshold of 4 mm; (d) log Odds Ratios for a threshold of 4 mm; (e) Skill scores for a threshold of 5%; (f) log Odds ratios for a threshold of 10%. The threshold of 5% is the lowest value of the highest 5% of pixels within the entire domain – the 95th percentile value. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. A skill score of 1.0 indicates a perfect forecast for a given square size; a score of 0.0 indicates no skill for a given square size. A log odds ratio of less than 0.0 indicates no skill, a value > ~1.0 is good. Further information on the skill score and odds ratios is given in the stage 4 report (Roberts 2004). Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

7.2.2 Mean accumulations

The key results from the graph in Figure 19 are:

1. The 1 and 4 km 'spin up' forecasts (dashed lines) started off producing too little rain and ended up producing far too much. The 1-km forecasts (red dashed) generated too much rain more quickly, but the 4-km forecasts (green dashed) 'overshot' by more.
2. The 1 km forecasts with added increments (red line) and the 4 km forecasts with data assimilation (green line) produced too much rain at the start, became comparable to the radar in the middle of the period and had slightly too little by the end (assuming zero error in the radar).
3. The 12-km forecasts (blue line) produced the closest mean accumulation to the radar.
4. The 4-km with data assimilation and the 12-km (also with data assimilation) were flat over the first two hours before dropping. This may be due to MOPS continuing past T+1.

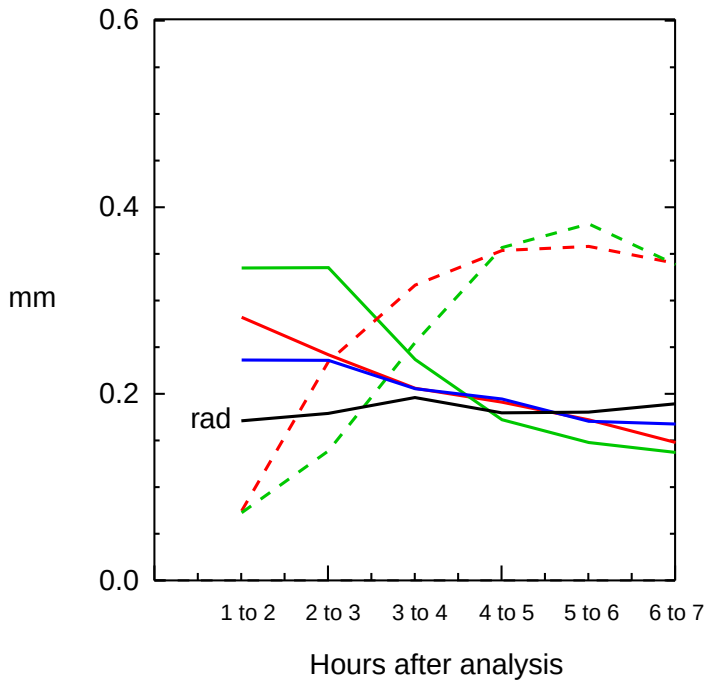


Figure 19. The mean hourly accumulations over the whole domain (in mm) for all four events. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. The black line = radar, blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

8 Conclusions

8.1 Summary of the results

- The fraction skill scores, Brier skill scores and (to a lesser extent) the odds ratios produced results that fit with the subjective assessment of the events. We can have confidence in these scores.

Mean rainfall amounts over the whole domain

- The 1 and 4 km forecasts that had to spin up from the 12 km T+1 forecasts produced too little rain at the start of the verification period and far too much at the end.

- The 1 km forecasts with added increments and the 4 km forecasts with data assimilation generated too much rain at the start of the forecasts and then settled down to produce only slightly too little by the end.

- The 12-km forecasts produced mean rainfall amounts that were closest to the radar.

Scores for 6-hour accumulations

- Skill increased with spatial scale as expected.

- The forecasts were mostly comparable for the threshold of 0.001 mm (essentially rain or no rain). The 12-km forecasts were slightly better.

- With an accumulation threshold of 4 mm, the 1 and 4 km 'spin up' forecasts were generally the best followed by the 1 and 4 km forecasts with data assimilation. The 12-km forecasts tended to be the worst because they failed to produce high enough accumulations (although they were the best for one event, 1st July 2003).

- The 1 and 4 km gridlength forecasts were very significantly better then the 12 km gridlength forecasts for the highest used accumulation threshold of 8 mm. With that threshold the 1 km gridlength forecasts with added increments were the best.

- The 1 and 4 km 'spin up' forecasts were significantly better than the rest when using a threshold of the top 10% of accumulations (90th percentile value). This indicates that the spatial distribution of the higher accumulations was better represented in those forecasts. The other forecasts were comparable to each other overall, though with variations from event to event.

- A useful spatial scale for the prediction of the location of the highest 6-hour accumulations was around 25 km for the 1 and 4 km 'spin up' forecasts, 55 km for the 1 and 4 km forecasts with added data and 70 km for the 12-km forecasts.

- Some individual forecasts could be very good. In particular, the 1 and 4 km 'spin up' forecasts for event A (28th Aug 2003) and the 1 km forecasts with added increments for event B (13th May 2003) stand out for a threshold of 4 mm.

Scores for hourly accumulations

- Hourly forecasts of rainfall accumulations were less skilful than 6-hour forecasts because timing errors became more important.

- The 12-km forecasts tended to lose skill during the forecast period and became poor by the end, particularly in two of the events. This was because showers decayed to quickly.

- The 1 and 4 km 'spin up' forecasts gave the worst scores in the first hour or two, especially for higher thresholds. The reason is that they had too little rain at the start. They were generally the best by the end of the period, although this varied from event to event.

8.2 Implications of the results

8.2.1 The impact of higher resolution

We can see that the 12 km forecasts did not produce high enough rainfall accumulations locally in these convective situations, although they did produce showers and rainbands in the right general area, (provided that the showers occurred during the day and over land).

The higher-resolution forecasts were able to generate the locally higher accumulations, though not necessarily in quite the right place. For an accumulation threshold of 8 mm (the highest appropriate for these events), the higher resolution forecasts gave the best scores (in particular, the 1 km gridlength with added increments). However, they also tended to produce too much rain.

So which is best? If we examine the ability of the forecasts to predict the spatial distribution of the higher accumulations, the results show that the 1 and 4 km 'spin up' forecasts were generally the most accurate and the 1 km forecasts produced the best results. Useful accuracy was achieved by these forecasts over spatial scales of around 20-25 km, compared to around 70 km for 12 km. This is encouraging because it suggests that higher-resolution models can spin up extra detail from the information contained within the 12-km model and generate more spatially accurate forecasts. It also implies that, for the most significant rainfall events, it would not be worthwhile to statistically post process a 12 km forecast to get more realistic rainfall accumulations, rather than run a higher resolution model, because if the distribution of the rain is poorer (over the spatial scale of interest) it remains poorer. All well and good, but we also know that the 1 and 4 km 'spin up' forecasts produced too much rain (though not enough at the start of a forecast). It does indicate that there are some problems representing the convection correctly. Instability that is not released at the start builds up and was then released in an unrealistic way. Part of this must be due to starting the forecasts from coarser resolution fields, but it also suggests that there may be difficulties in representing the convection correctly (and the two are linked). Research is in progress at the Joint Centre for Mesoscale Meteorology (JCMM) to examine the way turbulence in convection can be more realistically modelled and the way cloud microphysics can be better represented in high-resolution models.

8.2.2 The impact of data assimilation

The first point to make is that data assimilation is required. There is no point in running a storm-scale model from coarser resolution fields. The results from the hourly accumulations back this up. They show that the 1 and 4 km 'spin up' forecasts (with no high-resolution data assimilation) have low skill at the start, particularly for high accumulation thresholds. They were not useful for the first hour or two of the forecasts, and since we have an interest in the short-range prediction of high impact rainfall events, this is not satisfactory.

The data assimilation methods applied here (added increments at 1 km and 3DVAR and MOPS at 4 km) have had a beneficial effect in improving the skill at the start of the forecasts, but unfortunately this has not been sufficiently maintained throughout the forecasts. The over-prediction of rainfall amounts has been greatly improved later in the forecasts, but made worse at the start. The ability of the high-resolution forecasts with data assimilation to predict the spatial distribution of the higher accumulations is not as good as the high-resolution 'spin up' runs (though still better than 12 km). It appears that the data assimilation, by adding extra high-resolution information, was somehow disturbing the coherence in the development of convection that the 'spin up' runs had. It was, perhaps, making the model generate convection at the start of the forecasts in locations that were not compatible with the regions the model dynamics would have liked to generate them – and the compromise produced a poorer forecast. There is a indication that MOPS at 4 km had a detrimental effect on the forecasts of the highest accumulations (> 8mm).

We would like the best of all worlds; we want to run with high-resolution information from the start, have the data assimilation improve the skill at the start, and yet not interfere in a negative way with the larger-scale dynamics that may be crucial to the evolution of the forecasts. The data assimilation techniques examined here are only a first stab at achieving this. They have served to highlight the issues involved. Research is now in progress to find and develop data assimilation methods that are specifically designed for nested storm-scale model configurations, so that, hopefully, the 'best of all worlds' can eventually be achieved.

8.2.3 Final comments

These are encouraging results. They show quantitatively that improved resolution (down to 1-km gridlength) can have a positive impact and has the potential to significantly improve forecasts over the scales of small-medium sized river catchments. They also show that there is still plenty of work do for that potential to be fully realised.

It is acknowledged that this investigation has only examined a small number of events and that conclusions are only possible within that limitation. Further cases need to be examined, particularly of severe events that were lacking over the region of interest in the unusual summer of 2003.

9 Appendices

9.1 Appendix A. Data assimilation methods

9.1.1 3DVAR

3DVAR is the technique that is used to adjust the model to be closer to observations of pressure, temperature, humidity and wind. It is used operationally in the 12km-gridlength mesoscale model (and the global forecast model, see Lorenc et al 2000). It is now available to be tested in a model with a gridlength of 4km, but is currently too costly to run with a gridlength of 1km, and possibly, pointless.

The aim of 3DVAR is to minimise the misfit of a model state at a particular time to new observations and to a background state. The background state is a previous forecast. The model state is a reasonable guess to start with, which is usually the same as the background. The variables that define the model state are varied until a compromise is reached in which the combined misfit to both the observations and the background state is minimised. This process is crucially dependent on the expected errors in the background state and the observations and also on the influence that each model gridpoint is expected to have on every other. Once the new model state (the analysis) has been found the differences between this new analysis and the background state are found and stored as increments.

These increments are then added to the background model forecast bit by bit through a period of time. The reason for adding the increments in this gradual way is to enable the model variables to mutually adjust over a period so that noise is not generated by the shock. The time period used to add the increments from 3DVAR in the operational 12 km gridlength forecast system is 2 hours (see Figure 20).

The types of observations used are:

1. Surface reports of surface pressure, temperature, humidity and wind, from conventional surface stations, buoys and ships and wind profilers.
2. Radiosondes – that provide vertical profiles of temperature wind and relative humidity.
3. Aircraft measurements of temperature and wind.
4. Satellite measurements of winds by tracking cloud features in imagery
5. Satellite measurements of vertical profiles of temperature and humidity

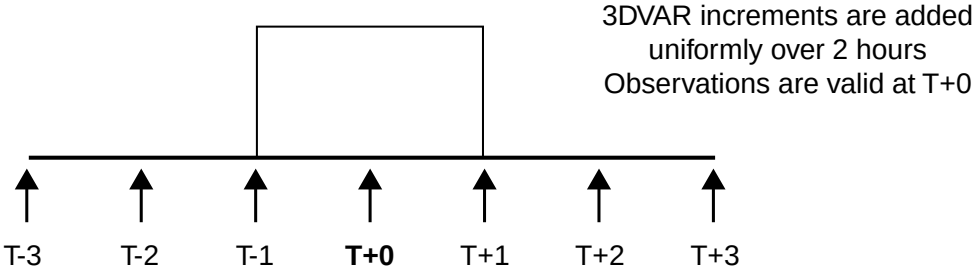


Figure 20. A schematic of the weighting given to observation increments added through 3DVAR. Courtesy of Zhihong Li.

9.1.2 Moisture Observation Pre-Processing System (MOPS)

There are two techniques used within MOPS :

1. Latent Heat Nudging
2. Cloud Analysis

MOPS Latent Heat Nudging

Latent Heat Nudging (LHN) (Jones and Macpherson1997) is a technique that has been developed to adjust the precipitation in a numerical forecast model to be closer to the observed precipitation. The method is used operationally in the 12km-gridlength mesoscale model. The rain rates are not modified directly. Instead, temperature increments are added to the model, to scale the model latent heating rate profile (heat that is generated from condensing water vapour into cloud and precipitation) by the ratio of observed rain rate to forecast rain rate. If the model is not producing rain where rain is observed, then a nearby point is used to supply a heating rate profile for scaling. The idea is, that once the temperature profiles have been adjusted, the model winds and vertical motion will adjust accordingly to produce a more realistic rainfall pattern (rain or snow results from saturated ascent).

The observed rain rates come from Nimrod (Golding 1998) processed radar data and are smoothed to the model grid if necessary. The weighting given to MOPS latent heat increments is shown in Figure 21.

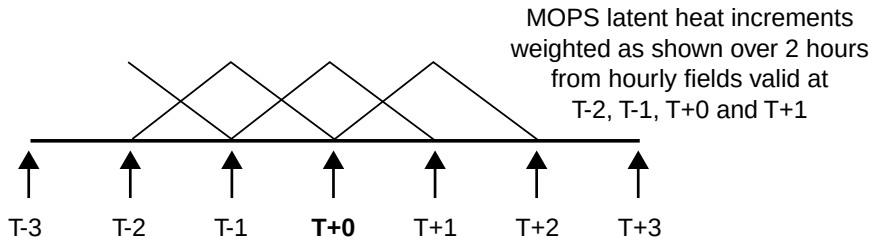


Figure 21. A schematic of the weighting given to observation increments added through latent heat nudging. Courtesy of Zhihong Li.

MOPS cloud analysis

This is a technique for adjusting the cloud in a model to give a better fit to the observed cloud structure (Macpherson et al 1996). It is used operationally in the 12km-gridlength mesoscale model. Nimrod radar analyses, surface rainfall measurements, Meteosat images and surface cloud cover reports are combined with the model background (a short forecast), to produce a 3-d picture of the 'actual' cloud amount in terms of cloud fractions within the model domain. The cloud fractions are then converted into profiles of relative humidity by use of the Smith scheme (Smith 1990) to relate cloud fraction to grid-box mean relative humidity. The humidity profiles are then nudged as increments into the model over a period of time. If both the MOPS cloud fraction and a model grid square have no cloud, then no nudging is done. Humidities are converted into cloud fraction within the model using the Smith scheme. The weighting given to MOPS latent heat increments is shown in Figure 22.

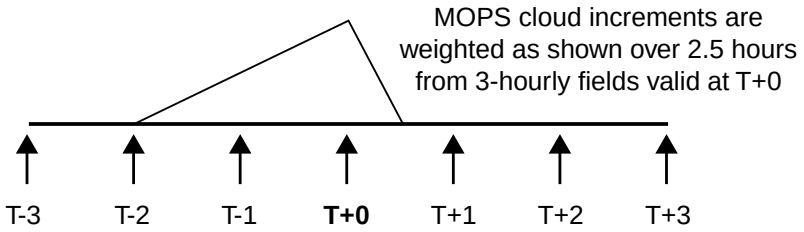


Figure 22. A schematic of the weighting given to observation increments added through latent heat nudging. Courtesy of Zhihong Li.

9.1.3 Addition of lower-resolution 3DVAR increments

Unlike the other methods, this approach is not used operationally. It is used specifically for high-resolution modelling because 3DVAR becomes increasingly computationally expensive as resolution is improved.

3DVAR is performed on either the 12 or 4 km gridlength model to produce a set of increments for a particular starting time. These increments are then added bit by bit to the higher resolution model we are interested in. For example, 3DVAR might usually be performed on a 4-km gridlength 3-hour forecast (called the 4-km background) to get a set of increments to add to that background in order to create a new 4-km analysis. Instead, these increments are added bit by bit to a 1-km gridlength background (an equivalent 1-km gridlength 3-hour forecast) in order to create a new 1-km analysis. The benefit of doing this is that a high-resolution forecast can be run without either having to start with fields interpolated from a coarser resolution model or run with 3DVAR at 1-km gridlength. It provides a way of preserving the high resolution detail from the previous forecast while adjusting the larger scales to match the observations.

9.2 Appendix B. Verification scores

9.2.1 Brier skill score

Brier score $BS = \frac{1}{N} \sum_{j=1}^{N} (p_j - o_j)^2$

$0 \leqslant p_j \leqslant 1$ forecast fractions (or probabilities) for a particular spatial scale (square size) and accumulation threshold

$o_j = 1$ if the accumulation threshold is exceeded or equalled

$o_j = 0$ if the accumulation is not exceeded

N = Total number of gridpoints

The Brier skill score (BSS) is given by

$$BSS = 1 - \frac{BS}{BS_{reference}}$$

Where $BS_{reference}$ is the reference Brier score given by

$$BS_{reference} = 2f_{ob}(1 - f_{ob})$$

It is the Brier score for a random forecast with the observed frequency f_{ob}

f_{ob} = the fraction of radar pixels exceeding the threshold over the whole domain

The BSS is only computed for $f_{ob} = f_{fc}$

where f_{fc} = the fraction of forecast pixels exceeding the threshold over the whole domain

E.g. We might compare the top 10% of radar accumulations with the top 10% of forecast accumulations then $f_{ob} = f_{fc} = 0.1$ (but the actual forecast and radar accumulation thresholds will probably be different)

Over the largest spatial scales, i.e. if the forecast fractions are taken over the whole domain then $p_j \longrightarrow f_{fc}$

$$BSS = 1 - \frac{f_{ob}(f_{fc} - 1)^2 + (1 - f_{ob}) f_{fc}^2}{2f_{ob}(1 - f_{ob})}$$

Then for $f_{ob} = f_{fc} = f$ over the largest spatial scales

$$BSS = 1 - \frac{f^3 - 2f^2 + f + f^2 - f^3}{2f(1 - f)} = 1 - \frac{f - f^2}{2(f - f^2)} = 0.5$$

Shown below is a graph of a typical curve for Brier skill scores from forecast fractions over increasing square sizes for a given frequency threshold (e.g. comparing the top 5% or 10% of forecast and radar rainfall accumulations over the whole domain).

The curve will always asymptote to 0.5 for forecast fractions computed over the largest spatial scales (width of squares = length of domain) - as long as a frequency rather than an actual rainfall accumulation or rate threshold is used.

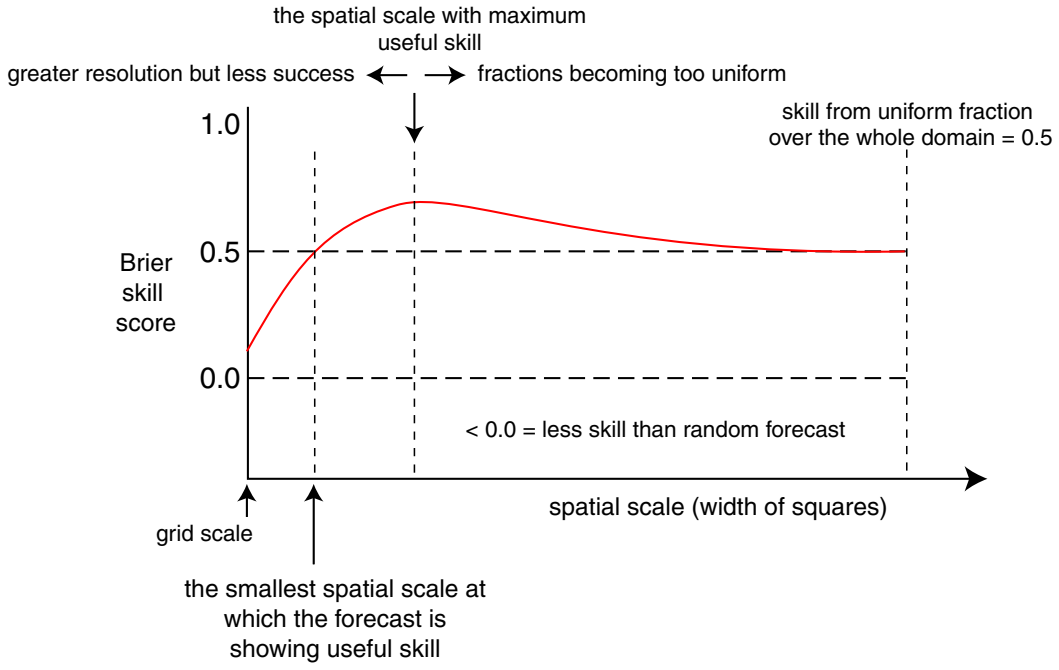
The score can have a maximum value of 1.0. This is only possible if the forecast and radar fields are an exact match, and can only occur at the grid scale. Such a scenario is extremely unlikely.

The worst possible score is determined by the frequency used for the threshold :-

Worst score = 1- (1/(1-f)) f = frequency threshold

The worst possible score can only be achieved at the grid scale. For a frequency threshold of f = 0.1 (the top 10% of observed and forecast accumulations), the worst possible score is -0.111. Any score less than zero indicates that the forecast has less skill than a random forecast.

The peak in the curve is a trade off between two competing effects. As the spatial scale (square size used to compute fractions) increases, more of the forecast fractions (for exceeding the threshold) will coincide with points where the threshold is exceeded. However, at some spatial scale, the incidence of forecast fractions coinciding with points where the threshold is not exceeded will become a stronger influence on the score. This happens as the fractions become more uniform (and less useful) with increasing spatial scale. It is the trade off between forecast resolution and reliability. If the curve does not go above 0.5, there will be no peak, indicating that the forecast has poor spatial skill. In this instance, the most skilful spatial scale is the size of the entire domain, which of course, is not very useful. The more skilful and useful the forecast, the smaller the spatial scale at which the peak occurs, and the sharper the peak will be. The spatial scale at which the curve crosses 0.5 is a useful quantity that can be detected automatically.



9.2.2 Fractions skill score

Fractions Brier score $BS_{fractions} = \frac{1}{N} \sum_{j=1}^{N} (p_j - o_j)^2$

$0 \leqslant p_j \leqslant 1$ forecast fractions for a particular spatial scale (square size) and accumulation threshold

$0 \leqslant o_j \leqslant 1$ radar fractions

N = Total number of gridpoints

The fractions skill score (FSS) is given by

$$FSS = 1 - \frac{BS_{fractions}}{\text{No Correlation Forecast}}$$

Where the 'No Correlation Forecast' is the $BS_{fractions}$ obtained if there is no correlation between the forecast and observed fractions, i.e. no overlap between any fractions > 0, and given by

No correlation forecast $NSF = \frac{1}{N} \left[\sum_{j=1}^{N} (p_j)^2 + \sum_{j=1}^{N} (o_j)^2 \right]$

Over the largest spatial scales, i.e. if the forecast fractions are taken over the whole domain then $p_j \longrightarrow f_{fc}$ and $o_j \longrightarrow f_{ob}$ for all points j

where f_{fc} = the forecast frequency (the fraction of forecast pixels exceeding the threshold over the whole domain)

f_{ob} = the radar frequency

$$\therefore \quad FSS = 1 - \frac{(f_{fc} - f_{ob})^2}{f_{fc}^2 + f_{ob}^2} = \frac{2 f_{fc} f_{ob}}{f_{fc}^2 + f_{ob}^2}$$

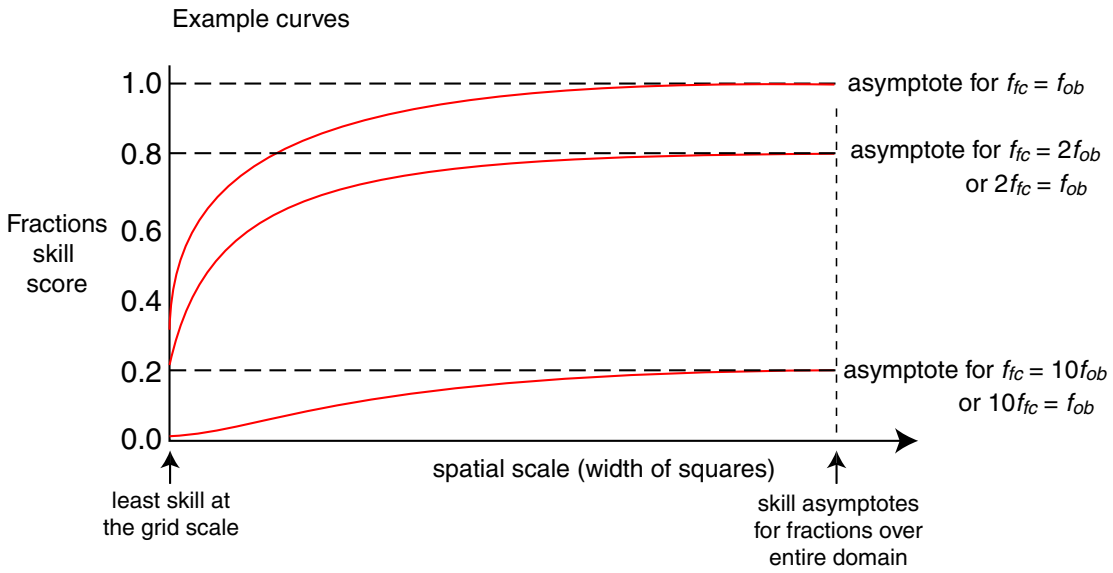
If either f_{fc} or f_{ob} = 0.0, FSS = 0.0, for all spatial scales

I.e. if no rain over the threshold is forecast but some is observed, the fractions skill score is always zero,
likewise, if rain over the threshold is forecast but none is observed, the fractions skill score is always zero

The score is dependent on the ratio of the fractions and is therefore more sensitive if either f_{fc} or f_{ob} is small. Hence, the score is more sensitive to forecast behaviour for rare events.

For fractions taken over the whole domain

If $f_{fc} = f_{ob}$, FSS = 1.0
If $f_{fc} = 2f_{ob}$ or $2f_{fc} = f_{ob}$ (one is twice the other), FSS = 0.8
If $f_{fc} = 10f_{ob}$ or $10f_{fc} = f_{ob}$ (one is ten times the other), FSS = 0.2



The graph shows some examples of curves that might be obtained from the use of the fractions skill score.

Skill improves or remains constant with fractions computed over increasing spatial scales. At the largest spatial scale (fractions computed over the whole domain), the observed and forecast fractions become the observed and forecast frequencies, and the score will asymptote to a value that is determined by those frequencies – as indicated in the graph. It does mean that, for rarer events, there is likely to be more scope for the observed and forecast frequencies to be significantly different – and therefore for the asymptotic score to be poorer. In other words, this score could tend to be harsher for rare events.

If either the observed frequency is zero or the forecast frequency is zero, the fractions skill score will be zero for all spatial scales.

9.2.3 (log) Odds ratio

For a given accumulation threshold, the 2x2 contingency table is

	radar >= fraction threshold	radar < fraction threshold
Forecast >= fraction threshold	A	B
Forecast < fraction threshold	C	D

A+B+C+D = N (the number of grid points)

The values of A, B, C and D are taken from the average of several contingency tables each using a different fraction threshold (say 0.1 (10%), 0.2, 0.3, 0.4 ..... 0.9)

$$\text{Odds ratio}, \theta = \frac{\text{Odds (hit)}}{\text{Odds (false alarm)}}$$

$$\text{Odds} = \frac{p(x)}{p(\bar{x})} = \frac{p}{1-p} \qquad p = \text{Probability of something}$$

$$\text{Probability of a hit} \quad H = \frac{A}{A+C}$$

$$\text{Probability of a false alarm} \quad F = \frac{B}{B+D}$$

$$\text{Odds (hit)} = \frac{A}{A+C}\left(1-\frac{A}{A+C}\right)^{-1} = \frac{A}{C}$$

$$\text{Odds (false alarm)} = \frac{B}{D}$$

$$\therefore \quad \text{Odds ratio}, \theta = \frac{\frac{H}{(1-H)}}{\frac{F}{(1-F)}} = \frac{AD}{BC}$$

Log θ is plotted rather than θ

If B or C are zero, θ can not be computed. If A, B, C or D are close to 0, the score can exhibit exteme behaviour and become unreliable. Large samples are required to overcome these shortcommings.

10 References

Golding, B.W. 1998: Nimrod: A system for generating automated very short range forecasts., Meteorol. Appl. 5, pp1-16

Jones, C.D. and Macpherson, B. 1997: A Latent Heat Nudging scheme for the Assimilation of Precipitation Data into an operational Mesoscale Model. *Meteorol. Appl.*, 4, 269-277.

Lean, H. W., 2003: High Resolution Trial Model project, results following the completion of stage 3. Available on request.

Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne and F. W. Saunders. 2000: The Met. Office Global 3-Dimensional Variational Data Assimilation Scheme. *Quart. J. Roy. Met. Soc.*, 126, 2991-3012.

Macpherson, B. 1996: The impact of MOPS moisture data in the UK Meteorological Office Mesoscale Data Assimilation Scheme., *Mon Wea Rev*, 124., 1746-1766

Roberts, N. M., 2004: Measuring the fit of rainfall analyses and forecasts to radar: Strategy for stage 5 evaluation. (Stage 4 report from the Storm-Scale numerical modelling project) *JCMM internal report.,* 146. *(NWP Technical Report.,* 432)

Smith, R.N.B., 1990. A scheme for predicting layer clouds and their water content in a general circulation model. *Q.J.R. Meterol. Soc.*, 116, 435-460

Stephenson, D.B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting*, 15, 221-232.